DELHI COLLEGE OF ENGINEERING

LIBRARY

Kashmiri Gate, Delhi-110006

Accession No.

Class No.

Book No.

Borrower is requisted to check the book and get the signatures on the torned pages, if any.

DELHI COLLEGE OF ENGINEERING
Kashmiri Gate, Delhi-110006
LIBRARY

DATE DUE

For each day's delay after the due date a fine of **10 Paise** per Vol. shall be charged for the first week, and **50 Paise** per Vol. per day for subsequent days.

Borrower's No.	Date Due	Borrower's No.	Date Due

ADVANCES IN ELECTRONICS

VOLUME III

ADVANCES IN ELECTRONICS

Edited by

L. MARTON

National Bureau of Standards, Washington, D. C.

Editorial Board

T. E. Allibone
H. B. G. Casimir
L. T. DeVore
W. G. Dow
A. O. C. Nier
W. B. Nottingham
E. R. Piore
M. Ponte
A. Rose
L. P. Smith

VOLUME III

1951

ACADEMIC PRESS INC., PUBLISHERS
NEW YORK, N. Y.

Copyright, 1951 by
ACADEMIC PRESS INC.
125 East 23rd Street
New York 10, N. Y.

All Rights Reserved

NO PART OF THIS BOOK MAY BE REPRODUCED
IN ANY FORM, BY PHOTOSTAT, MICROFILM,
OR ANY OTHER MEANS, WITHOUT WRITTEN
PERMISSION FROM THE PUBLISHER

PRINTED IN THE UNITED STATES OF AMERICA

CONTRIBUTORS TO VOLUME III

ASHWORTH, *Metropolitan-Vickers Electrical Company, Ltd., Manchester, England*

BLOCH, *Stanford University, Stanford, California*

BRILLOUIN, *International Business Machines Corporation, New York*

. CHODOROW, *Stanford University, Stanford, California*

L. GINZTON, *Stanford University, Stanford, California*

R. GUÉNARD, *Compagnie Générale de T.S.F., Paris, France*

A. GUILLEMIN, *Massachusetts Institute of Technology, Cambridge, Massachusetts*

EYER LEIFER, *Sylvania Electric Products Inc., Bayside, New York*

. F. MAYER, *School of Electrical Engineering, Cornell University, Ithaca, New York*

ILLIAM F. SCHREIBER, *Sylvania Electric Products, Inc., Bayside, New York*

USTAVE SHAPIRO, *National Bureau of Standards, Washington, D. C.*

. R. WARNECKE, *Compagnie Générale de T.S.F., Paris, France*

OHN E WHITE, *National Bureau of Standards, Washington, D. C.*

PREFACE

The present is the third of the series "Advances in Electronics." I and the members of the editorial board hope that it will be as favorably received as the preceding ones.

A volume of this kind is a product of the cooperation of many individuals. To the authors, editors, readers, publishers, and the many friends who contributed valuable ideas and advice, I acknowledge here my indebtedness and express my heartfelt thanks.

Some changes have again occurred in the membership of the editorial board of "Advances in Electronics." Dr. G. F. Metcalf in resigning recommended that he be replaced by Dr. Lloyd T. DeVore and we have been fortunate enough to secure Dr. DeVore's collaboration. I also have the pleasure to report that Dr. E. R. Piore, who in the past has helped us repeatedly with advice, consented to join the board.

L. MARTON

Washington, D. C.

CONTENTS

CONTRIBUTORS TO VOLUME III v
PREFACE vii

Field Emission Microscopy

By F. ASHWORTH, *Research Department, Metropolitan-Vickers Electrical Co. Ltd., Manchester, England*

I. Introduction 1
II. The Development of Field Emission Microscopy 2
III. Field Emission from Clean Metallic Surfaces 11
IV. Field Emission from Contaminated Surfaces 16
V. The Field Emission Microscope as a High Vacuum Gage 33
VI. The Resolving Power of the Field Emission Microscope 35
References 41

Velocity Modulated Tubes

By R. R. WARNECKE,* M. CHODOROW,† P. R. GUÉNARD,* AND E. L. GINZTON†

**Laboratoires de Recherches de la Compagnie Générale de T.S.F., Paris, France*
†Stanford University, Stanford, California

I. Introduction 43
II. The Basic Forms of the Klystron 45
III. Theory of the Klystron 50
IV. Klystron Amplifiers 66
V. Reflex Klystrons 71
VI. Summary 78
References 81

Electronic Theory of the Plane Magnetron

By L. BRILLOUIN, *International Business Machines Corporation, New York*

I. Introduction 85
II. Steady Case 89
III. Statement of the Problem. A Method of Integration Similar to Llewellyn's Procedure for a Diode 91
IV. Discussion of the Results. Standard Static Characteristic 92
V. Double Stream Solutions 97
VI. Transients and Oscillations Keeping the Plane Symmetry. Principle of the Method 99
VII. Operation of a Magnetron with a Short Impulse of Current 104
VIII. Discussion of British Reports on Similar Problems 108
IX. A General Discussion of Electron Trajectories in a Plane Magnetron 114

X. Steady Problem: Negative Resistance for Very Low Frequencies 116
XI. Small Oscillations of High Frequency: Fundamental Equations for the Trajectories . 117
XII. Characteristic Impedance of the Oscillating Plane Magnetron 119
XIII. Magnetron Impedance for Low Frequencies 121
XIV. Magnetron Impedance for High Frequencies 122
XV. Discussion of Some Special Examples 124
XVI. Double Stream Electronic Motions: General Formulas 128
XVII. Large Resonant Oscillations with Moderate Direct Current 132
XVIII. Efficiency and Negative Resistance in One-Anode Magnetrons 135
XIX. Physical Meaning of Conditions for Negative Resistance 139
Appendix 142
References 144

Electronic Theory of the Cylindrical Magnetron

By L. Brillouin and F. Bloch, *International Business Machines Corporation, New York, and Stanford University, Stanford, California*

I. Summary and Introduction 145
II. Basic Assumptions for Steady Conditions 147
III. The Problem's Equations. Static Case 148
IV. Fundamental Equation of Motion 150
V. Discussion of the Fundamental Equation 151
VI. Small Current, Small Oscillations 153
VII. Mathematical and Graphical Discussion of the Self-Consistent Trajectories 155
VIII. Standard Static Characteristics of Cylindrical Magnetrons 159
IX. Physical Interpretation 163
X. Cylindrical Magnetron under Variable Conditions 166
XI. Discussion of the Solution Obtained for Small Current 168
XII. Limits of Validity of the Single Stream Solution 171
XIII. Small Oscillations in a Cylindrical Magnetron 174
XIV. Calculation of the Anode Voltage 176
XV. Resistance and Reactance of a Cylindrical Magnetron 179
References 181

Tube Miniaturization

By John I. White, *National Bureau of Standards, Washington, D. C.*

I. Introduction 183
II. Limitations in Miniaturization of Tubes 184
III. Noteworthy Features of Subminiatures 189
IV. Summary State of the Art 192
References 194

Subminiaturization Techniques

By Gustave Shapiro, *National Bureau of Standards, Washington, D. C.*

I. Introduction 195
II. Design Philosophy 196
III. Thermal Considerations 197

IV Assembly Techniques 198
V Subminiature Assemblies 201
VI Components and Materials 208
VII Outstanding Problems 217
VIII Conclusions 218
References 218

Principles of Pulse Code Modulation

By H. F. Mayer, *School of Electrical Engineering, Cornell University, Ithaca, New York*

I Introduction 221
II Short Survey of Noise Cleaning Methods 222
III The Sampling Theorem 226
IV Quantization 229
V Encoding 233
VI Principal Operations at the Transmitting End 237
VII Principal Operations at the Receiving End 243
VIII Fidelity in PCM Transmission 247
IX Rate of Transmission 256
References 260

A Summary of Modern Methods of Network Synthesis

By E. A. Guillemin, *Massachusetts Institute of Technology, Cambridge, Massachusetts*

Introduction 261
I Analytic Form of an Impedance (Resp. Admittance) and Its Real Part 262
II Conditions and Tests for Positive Real Character 263
III Some Important Properties of Hurwitz Polynomials and Positive Real Functions 264
IV Special Forms of $Z(\lambda)$ in the Two-Element Cases 265
V Some Remarks Relevant to the Brune Process 267
VI The Darlington Procedure for the Solution of the Brune Problem Skeletonized 271
VII Synthesis of the Single-Loaded Lossless Coupling Network for a Prescribed Magnitude of Transfer Impedance 275
VIII Cauer's Method of Synthesis from a Specified $|Z(j\omega)|^2$ 276
IX Complementary Impedances; Constant-Resistance Filter Groups 279
X Another Way of Designing for Finite Resistances at Both Source and Load 281
XI The Constant-Resistance Lattice 283
XII An Alternate Realization Procedure for Transfer Functions 286
XIII Synthesis of a Lossless Two-Terminal-Pair Network through the Ladder Development of z_{22} 286

ILLUSTRATIVE EXAMPLES

XIV Brune's Synthesis Procedure 290
XV Darlington's Procedure Applied to the Same Problem 292
XVI An Alternative Method of Synthesis that Avoids Mutual Coupling 293
XVII Darlington's Procedure Applied to the Synthesis of a Transfer Impedance 295
XVIII Cauer's Method Applied to the Same Problem 296
XIX A Constant-Resistance Filter Group 297

XX. The Same Transfer Function Realized through a Lossless Network with Resistance Loading at Both Ends . . 298
XXI. Realization through a Cascade of Amplifier Stages 299
XXII. Further Illustration of the Ladder Development Procedure 300
References 303

Communication Theory

By Meyer Leifer and William F. Schreiber, *Sylvania Electric Products Inc., Bayside, New York*

I. Introduction . . 306
II. The Development of the Theory 307
III. The Synthesis of the Theory 320
IV. Applications to Television 339
References 343

Author Index 345
Subject Index 348

Field Emission Microscopy

F. ASHWORTH

Research Department, Metropolitan-Vickers Electrical Co. Ltd., Manchester, England

CONTENTS

Page
- I Introduction 1
- II The Development of Field Emission Microscopy 2
  - 1 The Cylindrical Form of Microscope 2
  - 2 The Spherical Form of Microscope 5
- III Field Emission from Clean Metallic Surfaces 11
- IV Field Emission from Contaminated Surfaces 16
  - 1. Experimental Results 16
  - 2. Theoretical Aspects of Surface Adsorption 16
  - 3. Factors Influencing Adsorption Processes 20
  - 4 Field Emission Studies of the Adsorption of Copper on Tungsten 27
  - 5. Field Emission Studies of More Complex Surface Processes 30
- V. The Field Emission Microscope as a High-Vacuum Gage . 33
- VI The Resolving Power of the Field Emission Microscope 35
  - 1. Observations on Individual Gas Atoms and Molecules 35
  - 2. Theoretical Aspects of Resolving Power 37
    - *a*. Particle Electrons 37
    - *b*. Wave Electrons 39
    - *c*. The Heisenberg Limiting Resolution 40
  - 3. Conclusions 40
- References 41

I. INTRODUCTION

Information about the structure of surfaces may be obtained from one or other of the electron optical techniques grouped together under the term "electron microscopy." Broadly speaking any instrument which utilizes electrons to produce a magnified image of a surface or section is an electron microscope. The conventional instrument is analogous to the optical microscope; electrons are transmitted through the specimen and focused. Other instruments focus or merely project electrons emitted by the specimen itself, a feature which restricts their use to the investigation of electron emitting surfaces. Theoretically any of the methods of inducing electron emission are available but only the thermionic,[1] photoelectric,[2] and field emission processes have been used to any extent. Few results have accrued from photoelectric studies using projection tubes because of experimental difficulties, and most information concerning surface properties has been obtained from thermionic and field emission work. The subject of thermionic emission microscopy has been reviewed up to 1949 by Herring and Nichols,[3] while the present article is devoted to field emission microscopy.

The field emission microscope is basically of simple construction. Nevertheless it requires both skill and patience to employ it successfully. The purpose of this article is to summarize the experience gained by those workers who have chosen to investigate its potentialities, and it will be shown that under suitable conditions this instrument may achieve resolving powers of atomic dimensions and magnifications up to 10^6. Surface phenomena such as the electron emission process itself and atomic and ionic adsorption may be intimately investigated.

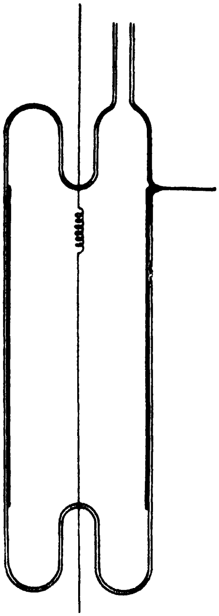


FIG 1 The cylindrical form of field emission microscope as used by Johnson and Shockley The emitting wire cathode is stretched along the axis of the cylindrical fluorescent screen anode.

II. THE DEVELOPMENT OF FIELD EMISSION MICROSCOPY

1. *The Cylindrical Form of Microscope*

In its simplest form the field emission microscope may consist of a tungsten filament as cathode stretched along the axis of an evacuated glass cylinder whose inner wall is coated with a semiconducting fluorescent material as the anode (Fig. 1). If several thousand volts are applied between the anode and the cold cathode, electrons are emitted from the latter under the influence of the high field intensity at the surface, which is a function of the applied field and the curvature of the surface. These electrons are emitted normally from the surface of the wire and travel radially across the tube to strike the fluorescent surface. The pattern of illumination on this screen is a complete picture of the state of the emitting surface. Axially the pattern is unmagnified but peripherally the magnification is as the ratio of the screen radius to the wire radius.

This type of microscope was first made by Johnson and Shockley[4] who used thin filaments of tungsten, molybdenum, tantalum, platinum, and iron as cathodes. They observed that the emission image on the cylindrical screen was determined by the condition of the emitting surface. In the case of the tungsten, for example, if the field was applied initially immediately after the experimental tube had been evacuated with the filament cold, an image consisting of bright and dark patches was observed. By raising the temperature of the filament these patches were modified and moved about, some vanishing permanently when the filament temperature was high enough. Patches were attributed either to field emission arising from regions of roughness on the surface of the filament or from impurities deposited on the surface during the drawing and handling of the wire. By raising the temperature sufficiently it was possible to remove permanently those patches due to contamination. When drawn wires were used as filaments, the electron images were streaked longitudinally by ragged bright lines arising from die marks on the wire surface. Since these marks obscured the essential features of the image, they were removed either by drastic heating *in vacuo* using alternating currents[5] until some of the surface metal had been evaporated or by rapid electrolytic etching. The smooth wire images resulting from this treatment were quite uniform in intensity, without streaks and sharp discontinuities, unless the individual crystals were so large that they were individually resolved. The size of the individual crystals in drawn wire is determined by the history of the wire and may be modified by suitable treatment.

Johnson and Shockley showed that the cylindrical type of microscope could be used to observe grain growth. Pure smooth tungsten wire was covered with cesium to enhance the emission. The cesium-on-tungsten image of the wire was uniform, indicating that the microcrystalline structure was smaller than the resolving power of the instrument. The outlines of individual crystals began to appear after the wire had been flashed at 3100°K, and the growth of these could be followed on the screen during subsequent heating. More detailed observations showed that the intensity of the electron emission was different for different tungsten crystal faces.

The resolving power of this arrangement was shown to be dependent on the dimensions of the system and the magnitude of the applied voltage. It was calculated by considering two electrons emitted from the surface, one normal to the surface along a radius with a velocity determined only by the applied field, the other having an additional thermal velocity normal to the radius. If a and b are the radii of the wire and screen respectively and the thermal velocity of the second electron is $(2kT/m)^{1/2}$, the angular separation between the points where these electrons strike the screen is given in radians by

$$\theta = 2\left[\frac{300kT}{ev}\right]^{1/2}\left[\log_e \cdot \frac{b}{a}\right]^{1/2}\int_0^{(\log_e b/a)^{1/2}} e^{-u^2}du \qquad (1)$$

For a potential of 10 kv between cathode and screen the peripheral resolving power would therefore be 1° for a 0.0025-in. radius filament with a 2-in. radius screen. The magnification of this system is 2000 and represents a reasonable order for this type of emission microscope. A limiting magnification of several thousand is dictated by the minimum radius of drawn wire which can be used as the filament.

A number of workers have used cylindrical type microscopes to observe wire surfaces activated by material issuing from the interior. The work of Brüche and Mahl[6] with thoriated molybdenum and of Ahearn and Becker[7] with thoriated tungsten, helped to clarify ideas about the process of activation. Even after intense heating, cathodes of these materials produced images covered with bright patches. When the temperature was increased these patches spread into rings and simultaneously other spots appeared and began to spread, until ultimately the surface appeared to be covered completely by the activating material. The phenomenon suggested that the activating material arrived at the surface by random eruptions from the interior, and it appeared that these eruptions did not necessarily occur preferentially at crystalline boundaries.

All surface studies by cylindrical emission microscope before 1940 had used polycrystalline wires, and this made the interpretation of results difficult. In spite of this it had been established that the intensity of emission from both pure and activated tungsten surfaces depended on their crystallographic form. This was confirmed when Nichols[8] showed how to produce long single crystals in pure tungsten wires and used selected specimens in a cylindrical microscope to measure both thermionic and field emission from particular crystal faces. The method was based on the Pintsch process of passing polycrystalline tungsten wire, with suitable fluxing agents, through a steep temperature gradient. Pintsch specimens may be meters long and fill the whole cross section of the wire, but unfortunately they contain impurities from the flux. Nichols started with a pure tungsten wire already containing a seed crystal, thus eliminating the need for flux, and extended this seed to fill the whole cross section by passing a steep temperature gradient along the wire.

Crystal growth was also studied by Robinson[9] in 1941. Pure tungsten wires were ground uniformly by a method described earlier[10] until a high-powered microscope showed only minute scratches. The wires were used in the microscope after flashing at 3000°K to remove surface contaminations. The electron emission patterns then showed the absence of die marks and indicated a fine-grained polycrystalline structure. The growth of single crystals by heat treatment was then studied at temperatures between 1900°K and 2200°K. In some wires only one crystal was involved in the growth whereas in others, a number of crystals grew in different parts of the wire eventually merging together. Growth diagrams were plotted for number of wires and an exponential rate of growth found which could be described by the expression

$$R = A \cdot \exp(-b/T) \text{ cm/hr} \tag{2}$$

with a value for b of (55,000 ± 19,000) °K. At 1900°K the time for recrystallization is of the order of a hundred hours while at 2200°K it is only five minutes.

The cylindrical form of emission microscope suffers from a number of disadvantages, chiefly the anisotropic magnification and resolution which make interpretation of the image rather difficult. The minimum radius of the specimen wire defines the maximum magnification of the system which is certainly not greater than 10^4. It will be shown in the next section that the spherical form of the emission microscope offers isotropic magnification and resolution making interpretation relatively easier, with magnifications up to 10^6 times and resolving powers of atomic dimensions.

2. *The Spherical Form of Microscope*

In 1936, Müller[11] constructed a field emission microscope consisting of a microcrystal of tungsten mounted at the center of curvature of a concave fluorescent screen (Fig. 2). The microcrystalline specimen was prepared by etching the end of a fine wire of the metal to be investigated. Tungsten of 0.05-mm radius was etched in molten sodium nitrite to produce an end radius of 1 micron. The specimen wire was spot welded to the tip of a hairpin heating filament of thicker tungsten wire to support it, as shown. By applying several kilovolts between the point specimen and the screen, field emission images of the surface were produced with magnifications of the order of 10^5.

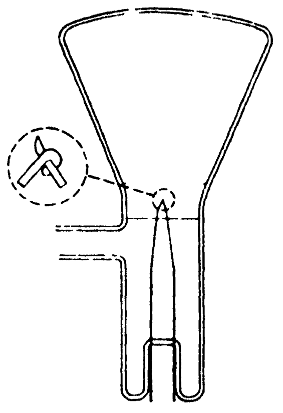


FIG. 2. The spherical form of field emission microscope as used by E. W. Muller. The emitting tungsten crystal is situated at the end of the etched tungsten wire (inset) mounted as shown in the form of a hairpin heating filament. The fluorescent screen anode covers the end face of the tube.

Müller suggested that the technique could be used for studying the properties of metallic surfaces and adsorbed layers. He published image pattern photographs of the field emission from tungsten (Fig. 3a)[27] and molybdenum (body-centered cubic structures), copper and nickel (face-centered cubic structures), and tungsten contaminated with oxygen. The symmetry of the images obtained with specimens whose surfaces could be thoroughly cleaned by heating bore a direct relationship with their crystallographic symmetry (Fig. 3*b*). The images for both clean tungsten and molybdenum were found to be similar and their symmetry could be simply associated with their body-centered cubic structure. The copper and nickel specimens could not be adequately cleaned on account of their low melting points; their emission images were unstable and not reproducible. Müller illustrated how a tungsten emission pattern was modified by admitting oxygen to the system, thus contaminating the surface of the specimen. Preferential poisoning occurred over the region surrounding the (100) face, the emission from this region being suppressed. By raising the temperature of the specimen for short periods and then observing the emission patterns from the cold surface, the gradual removal of oxygen from different parts of the surface was followed until after heating to 2200°K it appeared that all the oxygen had been driven off and the original tungsten pattern was left. To produce

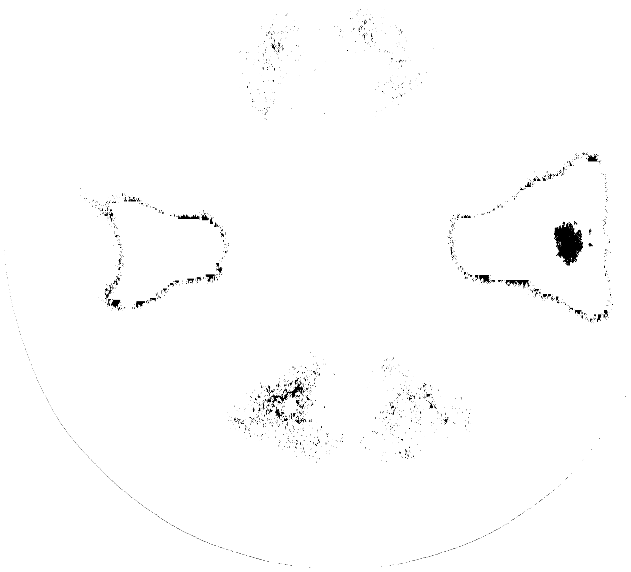


FIG. 3*a* Field emission image for clean tungsten "built-up" by the application of field to the surface during flashing

the same order of emission intensity an oxygen-contaminated surface required a higher field stress than a clean surface.

More detailed work followed[12] in which observations were made on the mobility of tungsten atoms over tungsten surfaces, and it was claimed that individual atomic migration could be detected when the surface temperature was increased to about 1100°K Certain preferred directions of flow were thought to be defined by the surface lattice structure.

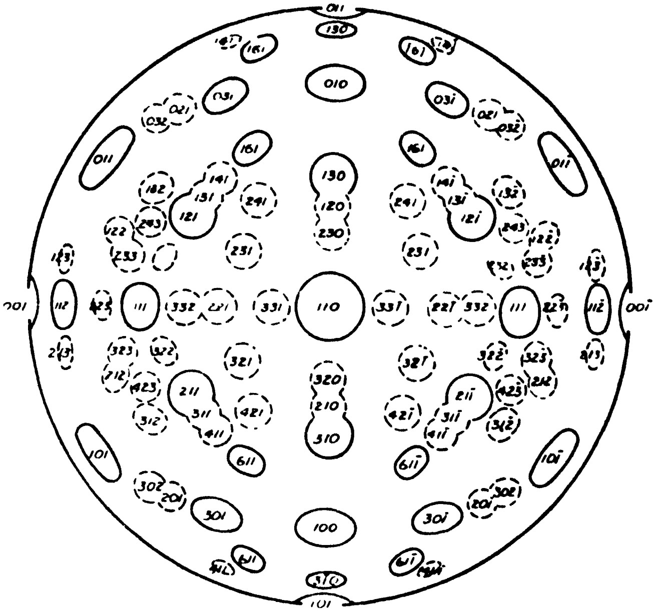


FIG. 3b. The stereoscopic projection of the crystal axes (in terms of their Miller indices) of a body-centered cubic structure e.g., tungsten. Full-line circles correspond to those regions of the tungsten field emission pattern which are readily identified with crystal faces. Dotted circles correspond to those regions not so readily identified.

What was considered to be clean tungsten displayed emission from all regions except those corresponding by comparison with the crystallographic symmetry, to the (211), (100), and (110) faces. The nonemitting regions were thought to be either depressions in the approximately hemispherical surface or localities having comparatively high work functions.

The behavior of barium deposited on a tungsten surface was described in some detail. A thick layer was first deposited and then the surface heated. Crystallites appeared to have grown after heating to 800°K, and at higher temperatures these split up and diminished in size, suggesting that barium was lost by evaporation. When the temperature was raised to 1500°K for a few seconds, the deposit was reduced to a monatomic layer on the tungsten surface, and Müller claimed that individual ions could be observed as moving bright spots. When a thin layer of barium was deposited, migration over the tungsten surface began at 690°K. More recent work has shown similar behavior of deposits of other materials which will be described later.

Müller's experiments indicated some advantages of the field emission projection technique over the corresponding thermionic emission observations,[13] particularly since they permitted surface processes to be observed at room temperature.

The apparent correlation between emission pattern and crystal symmetry led Müller[14] to suggest that the nonemission directions of field emission corresponded to the nontransmission directions of x-rays through that type of crystal lattice. The theory due to W. L. Bragg of the transmission of x-rays through crystals expresses the relationship between the wavelengths of the x-ray beam λ and the lattice parameter *a*, as follows

$$n\lambda/2 = ma/(h^2 + k^2 + l^2) \tag{3}$$

where *m* and *n* are integers and (*h*,*k*,*l*) are the Miller indices of crystal lattice direction.

For a body-centered cubic lattice, *e.g.*, molybdenum and tungsten, x-ray reflections at lattice planes will occur for particular wavelengths, for those directions whose Miller indices add up to even whole numbers. Similarly for a face-centered cubic lattice, e.g., copper and nickel, the conditions for nontransmission are that the Miller indices are all odd or all even. These considerations alone, however, do not describe the experimental results and other factors must be taken into account. The wavelengths λ of the conduction electrons in metallic crystals are related to their energies (electron volts) by the de Broglie relation

$$\lambda = (150/\text{ev})^{1/2}\ \text{Å} \tag{4}$$

For tungsten the calculated conduction electron energy values are 5.8 ev assuming an electron population of one per atom and 9.2 ev for two per atom. A nominal value of 7.5 ev gave Müller an electron wavelength of 4.5 Å. Using this he obtained reflections for tungsten for example for the (211) direction when $m = 7$ and $n = 5$ and for the (100) direction when $m = 5$ and $n = 7$. His hypothesis did not explain the intensity distributions of the emission pattern.

The next attempts to understand this phenomenon were made by Benjamin and Jenkins[15] who used apparatus similar to Müller's and investigated both single crystals and polycrystalline specimens of tungsten, molybdenum, and nickel. The tungsten and molybdenum specimens were initially etched electrolytically in a 10% solution of sodium hydroxide and finally polished in a 10% solution of ammonia. The nickel specimens were etched in a solution of potassium perchlorate in hydrochloric acid. By observing the emission images the movements of atoms in the surface of the specimen could be followed, and the onset of surface mobility was found to occur at 1170°K for tungsten, 770°K for molybdenum, and 370°K for nickel. If the field intensity was maintained and the temperature suddenly reduced, the transient image could be frozen to permit detailed observation. At temperatures greater than the onset values, surface atoms were seen to flow preferentially in certain directions and particularly toward certain crystal edges which could be built up by simultaneous application of field and temperature. If the temperature was maintained with the field switched off the built-up regions gradually dispersed.

The electron images obtained by Benjamin and Jenkins for clean metallic surfaces were similar to those obtained by Müller, and a further attempt to explain the intensity distribution was made. The Miller indices corresponding to several crystal directions for which the field emission intensity was suppressed were determined and the electron wavelengths and energies which would give first order reflections for these directions were calculated. Second and higher order reflections would require electrons of shorter wavelengths and higher energies. Table I summarizes their results. The Fermi energy for the conduction electrons in tungsten has been determined[16] for two cases. If it is assumed that there is only one conduction electron per atom, the value is 5.8 ev; if two electrons per atom, it is 9.2 ev. For nickel, assuming two electrons per atom, it is 11.7 ev. Only for the (110) direction for tungsten and for the (111) direction for nickel are electrons of the correct order of

TABLE I

Crystal Structure	Nonemitting Directions	First Order Reflection Electron Energy
Body-centered cubic	110	7 48 ev
	200	13 98
(molybdenum and tungsten)	211	22 44
	310	37 40
Face-centered cubic	111	9 06
	200	12 08
(nickel and copper)	220	24 16
	311	33 22

energies available for total internal reflection. This approach failed to *explain the observed emission images. Nevertheless the correlation* between the symmetry of the emission pattern and the known crystal symmetry was apparent.

Müller's experiments were continued by Haefer[17] who used a transmission type electron microscope to study the size and form of his crystal specimen. He summarized his observations as follows:

(1) The surface of an etched tungsten point is not necessarily smooth but may consist of a collection of microcrystalline points.

(2) The surface roughness on an etched tungsten point may be removed by heating to 2400°K for thirty minutes. Even after this treatment, however, tiny crystal outcrops with radii of the order of 0.1 micron may remain on the smoothed surface.

(3) Final surface polishing to remove the tiny outcrops may be carried out by using the surface as a field emitter. Most of the emission will occur from these outcrops around which a high field intensity will develop. By drawing large emission currents from these centers they become heated until in the molten state they flow smoothly over the surrounding surface.

(4) The radius of curvature of a crystal point may be reduced by bombarding it with argon ions. An ion current of 10^{-6} ampere in argon at a pressure of 10^{-5} mm Hg is useful for this purpose.

Haefer observed the migration of barium, potassium, and cesium atoms over tungsten crystal surfaces and later determined the work functions of surfaces completely and partially covered, by substituting his experimental values of applied field, emission current, and emitting area into the Fowler-Nordheim image field corrected expression for field emission.[18] The field emission microscope permitted the determination of the true emitting area in these measurements and enhanced the value

TABLE II

Cathode Surface Condition	Applied Potential	Work Function
Clean tungsten point	5 2 kv	4 5 ev
Small quantity of barium. (almost all surface emitting)	4 86	4 31
Emission only from discrete patches of barium as more is deposited	3 82 to 1 8	3 26 to 1 87
Barium now lying on side of surface remote from source	1 55	1 67
Compact barium layer with some crystallites:	2 3	2 45
Evaporation of the barium from the point gave for the optimum thickness conditions	1 4	1 63

of the results. Work functions for thick layers of adsorbed alkali metal ions and optimum values for thin layers were thus obtained, as summarized in Table II.

The adsorption, migration, and evaporation of barium, thorium, and sodium on molybdenum and tungsten were studied by Benjamin and Jenkins using the field emission microscope and a number of photographs were published[19] but no detailed analysis of the results was presented.

The alkali and alkaline earth metals have received most attention in the form of adsorbed layers on molybdenum and tungsten because they are fairly simple materials for the vacuum technician to handle More recent observations made by the author on the behavior of copper atoms on tungsten are discussed in Sections IV, 4 and IV, 5.

The foregoing account must serve as an introduction to the subject while the most recent work is considered in more detail in Sections III and IV which deal with the emission from clean surfaces and from adsorbed layers.

III. FIELD EMISSION FROM CLEAN METALLIC SURFACES

Very little has been published concerning detailed observations of the pure field emission from clean metallic surfaces. Clean tungsten has been studied by Müller,[11,12,14] Benjamin and Jenkins,[15] and the author (1947, unpublished). Molybdenum,[15] nickel,[11,15] and copper[11] have also received attention, but the results are doubtful on account of the difficulties of removing adsorbed contaminations.

It is established that the field emission intensities for a given applied field are different for different crystal faces of a clean metallic surface. This suggests that different faces have different work functions, and in fact photoelectric[20] and thermionic[8] emission measurements seem to confirm this. A number of hypotheses proposed to explain these results will be briefly discussed:

(1) Smoluchowski[21] suggested that the surface work function might be the sum of one contribution due to the bulk of crystal beneath the surface—a volume effect—and another due to the effective double layer at the surface (see Wigner and Bardeen[22]). From the crystal lattice surface structure he determined the differences to be expected between the work functions of different crystal faces. The surface contribution was found by distributing the positive charge of each ion throughout its appropriate atomic cell while the negative charge of the conduction electrons was spread uniformly throughout the whole lattice. The resultant charge was then zero within the crystal. At the surface the boundary of the conduction electron cloud was smoothed out while the positive charge distribution remained unsmoothed. In this way the

resultant crystal surface became divided into regions of different charges. For a simple cubic lattice some crystal faces, such as the cubic face (100), *would have a dipole layer with the negative charge outside the surface, others, such as the rhombic dodecahedral face (110), would have the* positive outside the surface. To develop these ideas further and obtain numerical results, Smoluchowski made some simplifying assumptions. The calculated values of work functions did not agree with those obtained by Nichols[8] by the thermionic method although the values for the (100), (110), (111), and (112) faces had the same sequence of decreasing magnitude.

(2) Muller[14] and Benjamin and Jenkins[15] considered the possibility that the anisotropic Fermi surface inside the crystal could lead to different work functions for different crystal directions. The conditions for electron transmission through the crystal have already been discussed in terms of Bragg's law in Section II, 2 where it was shown that impossibly high values for the energy of the conduction electron were required to explain the observations.

(3) The theory proposed by Stranski and Suhrmann[23] to account for the apparently different work functions for different crystal faces suggested that one or more of three factors could be responsible. These were the specific surface energy, the "separation work" of the loosest surface component, and the shortest distance between adjacent surface components. Their results did not agree with Nichols' observations [8]

(4) A recent qualitative theory by the author considers the field emission intensity differences between crystal faces in terms of the atomic structure of the surface. Two factors may determine the relative field emission intensities from particular faces, the relative number of emitting centers and the magnitude of the local field intensity in the neighborhood of these centers. The qualitative analysis of a field emission image for a "built-up" clean tungsten microcrystal surface (Fig. 3*a*) in terms of a crystal surface diagram (Fig. 3*b*)* permits the faces to be listed in the order of increasing emission intensity as shown in Table III.

TABLE III

a	*b*	*c*	*d*	*e*	*f*	*g*	*h*	*i*	*j*	*k*
(011)	(112)	(001)	(113)	(013)	(111)	(012)	(116)	(023)	(122)	(233)

Assuming the crystal specimen in question to have a pseudospherical surface (produced by etching), it is purely a matter of spherical trigonometry to construct a model whose faces correspond to crystal faces. The size of each face is determined by the limitation that not more than *one surface atom of the next layer shall be enclosed within the true sphere. On this basis an approximate correlation is obtained between* the relative sizes of faces and of their observed emission images.

---

* There is a unique correlation between the stereoscopic projection of the crystal axes (Fig. 3*b*) and the observed emission pattern. The central dark (nonemitting) spot of Fig. 3*a* is identified as arising from the (110) crystal face by the particular symmetry of the surrounding emission pattern.

A detailed knowledge of the crystal structure in the neighborhood of the metal surface can be obtained geometrically if one assumes that the crystal structure of the body of the metal persists without modification

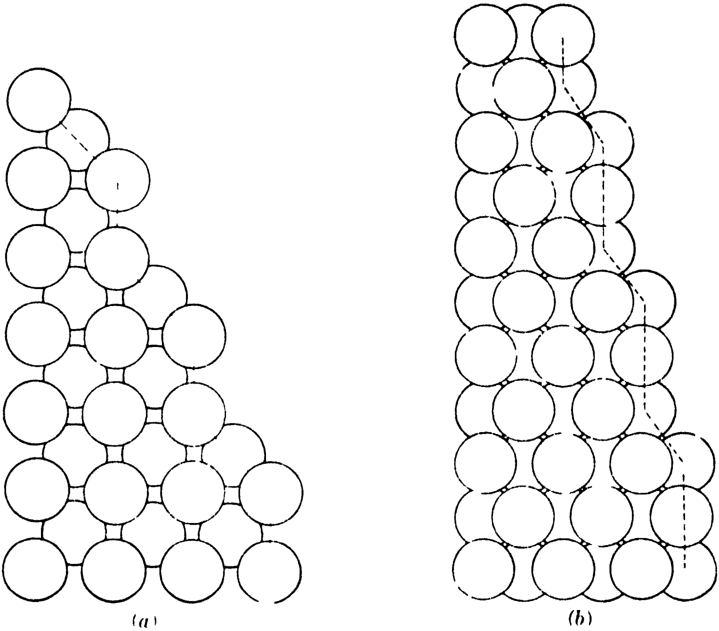


FIG. 4. Sections through typical body-centered cubic crystal faces; (*a*) the (012) face and (*b*) the (113) face.

right up to surface. Each crystal face will then have its own characteristic structure as shown, for example, in the profile views of the (012) and (113) faces for a body-centered cubic crystal (Fig. 4). Each face consists of a characteristic hill and valley structure. Some faces are rougher in this sense than others, and the atoms are packed together in different degrees of closeness for different faces. If we consider only those atoms which are situated in the surface, we may distinguish between two extreme types, those which are situated deepest in the valleys and those which occupy the most exposed sites in the hilltops. Some idea of the relative numbers of neighbors of different orders which are associated *with atoms in these two types of site for several faces of a body-centered* cubic crystal are given in Table IV. For a particular site there will be:

n_1 nearest neighbor atoms
n_2 next-nearest neighbors
n_3 next-next-nearest neighbors
n_4 next-next-next-nearest neighbors

The distances between atom centers for these different orders of neighborliness for a body-centered cubic crystal with a lattice constant of a Å, are r_1, r_2, r_3. and r_4 respectively, where

$$r_1 = a\sqrt{3}/2; \quad r_2 = a$$
$$r_3 = a\sqrt{2}; \quad r_4 = a\sqrt{11}\ 2$$

Stranski and Suhrmann[23] suggested that the relative emission intensities from different crystal faces might be correlated with the relative surface site binding energies This could be so if the site binding energy can be considered to give a measure of the binding energy of the conduction electrons in the surface which contribute to the local emission. The site binding energy will be a function of the numbers of the various neighbors as listed in Table IV However, this consideration alone cannot account for the relative order of emission intensities listed in Table III

TABLE IV

Crystal Face	Hill Site n_1	n_2	n_3	n_4	Valley Site n_1	n_2	n_3	n_4	No of Sites per sq cm Hills	Flats	Minimum Hill Angle
									($\times 10^{14}$)		
011	5	3	6	12	6	4	7	14	7 1	7 1	180 flats
112	5	3	6	12	5	3	7	11	4 75	4 75	180—flats
001	4	4	6	12	4	5	8	12	5 0	5 0	180 flats
113	5	3	6	12	6	4	7	14	6 5	6 5	150 –ridges
013	4	3	7	12	4	4	6	14	5 85	11 7	135 ridges
111	4	3	7	12	4	3	9	12	8 2	8 2	135 ridges
012	4	4	6	12	4	4	6	12	4 4	8 8	135 - ridges
116	4	3	7	12	4	3	7	12	1 6	9 6	135 corners
023	4	4	6	12	4	4	6	12	3 6	14 4	135 corners
122	4	3	7	12	4	3	7	12	5 75	11 5	135 corners
233	5	3	6	12	4	3	7	13	3 4	13 6	120—corners

Note. The crystal faces are arranged from top to bottom in order of increasing emission intensity. All atoms *within* the crystal have the following neighbors:

$$n_1 = 8, \; n_2 = 6, \quad n_3 = 12, \quad n_4 = 24$$

Two other factors which can influence the emission will be discussed: (1) the population density of the emitting sites and (2) the local field intensity

(1) The electron emission from a crystal surface must arise from the individual emitting sites, distributed uniformly or randomly over that surface. In the case of a clean crystal face we may assume that the distribution of these sites is uniform and a function of the repeating pattern of the surface structure. It follows that some faces will have a greater emission site population density than the rest and, other things being equal, those faces will yield relatively greater emission intensities. The population densities for sites in a number of faces are given in Table IV.

(2) Field emission is a function of the surface field intensity which is influenced by the radius of curvature of the surface. It is feasible to assume that the field intensities in the vicinity of surface projections of atomic dimensions will be sensitive to small differences in effective radii of curvature. From our knowledge of the surface structure it is possible to describe an emission site in terms of the sharpest surface projecting angle which occurs at that site. The approximate minimum angles subtended by the rows of atoms which meet at edges or corners in the surface (see Fig. 4) are given in Table IV.

The principal crystal faces are listed in Table IV in order of increasing emission intensity for a clean tungsten surface. Neither the relative site binding energies as indicated by the numbers of neighbors nor the relative numbers of sites per unit area have any direct correlation with the relative order of emission intensities. The order of decreasing minimum hill angle, however, follows the order of increasing emission intensity. More work is required to establish the relative importance of the various factors cited above in determining the relative emission intensities, but it is clear that the sharpness of the projections which form an integral part of some crystal face structures is one influencing factor. These structural corners and edges are the centers of high surface field intensities and the origin of preferential electron emission.

Although the above discussion is confined to the relative field emission intensities from different crystal faces, the relative values of apparent work function for these faces may be considered in the same terms. Faces for which the emission intensities are found to be small may be considered to have high work functions and vice versa. It is natural to ask whether the lists of work functions (or emission intensities), arranged in order of decreasing magnitude for the different faces, are the same, regardless of the emission process involved. Actually the thermionic emission experiments of Martin[13] and Nichols[8] suggest orders different from those observed by field emission. It may well be that the true or intrinsic work function of all faces of a given crystal surface is the same but that the differences in measured values (or observed emission intensities) are due to different disturbance factors introduced by the circumstances of measurement. In the case of field emission, the electrons are predominantly influenced by the local high surface field intensities; in thermionic emission the local fields may not be as important as the fact that the surface structure itself is in a thermally excited state, whereas in photoelectric and secondary emission, the relative ease with which the incident radiations or electrons can penetrate the crystal faces may be important.

At the present time we can only speculate in the absence of adequate experimental evidence. For results to be of any value the experiments must be carried out on absolutely clean surfaces, and one cannot be sure that all published work is reliable in this respect.

IV. FIELD EMISSION FROM CONTAMINATED SURFACES

1. *Experimental Results*

Both thermionic and field emission experiments have provided evidence of nonuniform adsorption of foreign atoms over the surfaces of metallic crystals. The relevant results obtained by projection tube studies for tungsten and molybdenum surfaces are presented in Table V and are discussed in this section. A detailed discussion of entry (15) in this table appears in Section IV, 4.

2. *Theoretical Aspects of Surface Adsorption*

Early theoretical work on the properties of metallic crystal surfaces treated the surface as a plane bounding a region of uniformly distributed negative and positive charges. Although the structure of the crystal was known, the surface structure was neglected. Adsorbed layers were treated as if they consisted of patches of ionized atoms or polarized molecules, but their detailed structure was not considered. Plausible hypotheses were presented to explain experimental results, and many of the most important of these have been collected by de Boer.[24] More recently, thermionic and field emission experiments by a number of workers cited in Table V, and, in particular, measurements of adsorption coefficients by Roberts and others[25] have suggested that surface properties can only be described in terms of the crystal structure of the surface. Furthermore, the assumption that contaminations are adsorbed in the ionic form was based on the inconclusive evidence that when contaminated surfaces are heated to high temperature, positive ions are emitted. From considerations of atomic and ionic sizes and packing requirements for close fitting between adsorbed layers and adsorbent structures, Stranski and Suhrmann[26] have deduced that cesium and

TABLE V

	Metal Surface	Adsorbed Atoms	Observer	Least Emission from	Adsorbed Atoms Prefer	Migration Temperature of Adsorbed Atoms	Evaporation Temperature of Adsorbed Atoms	Work Function in Electron Volts
(1)	Tungsten b c c $a = 3.16$ Å $d = 2.74$ Å	Oxygen $d = 1.2$ Å	Muller[11, 12] Haefer[17]	(100)	(100)		Completely evaporated at 2200°K	
(2)	Tungsten (see 1)	Oxygen (see 1)	Benjamin & Jenkins[1]				Evaporation commences at about 650°K	
(3)	Tungsten (see 1)	Barium b c c $a = 5.02$ Å $d = 4.2$ Å	Muller[14]	110 112 130 100	Between 221 & (111) 120	Starts at 690 K for less than a monatomic layer Tends to form crystallites	Starts at 1500°K but leaves small groups of atoms strongly bound at this temperature	
(4)	Tungsten (see 1)	Barium (see 3	Haefer[1]					2.5 for very thick layers 1.56 for optimum thickness
(5)	Tungsten (see 1) and Molybdenum b c c $a = 3.14$ Å $d = 2.72$ Å	Barium (see 3	Benjamin & Jenkins[16]	110 100	(211	Starts at 770 K	Starts at 900 K from (110) and (100) first Bound most strongly to 211	

TABLE V.—(*Continued*)

	Metal Surface	Adsorbed Atoms	Observer	Least Emission from	Adsorbed Atoms Prefer	Migration Temperature of Adsorbed Atoms	Evaporation Temperature of Adsorbed Atoms	Work Function in Electron Volts
(6)	**Tungsten** (see 1)	Cesium b.c.c. $a = 6.2$ Å $d = 5.4$ Å	Martin[13] (thermionic emission)		(110); (211). Boundary of (111). "Atoms are most strongly bound on faces having high work functions"			
(7)	**Tungsten** (see 1)	Cesium (see 6)	Haefer[17]					1.93 for very thick layers 1.36 for optimum thickness
(8)	**Tungsten** (see 1)	Cesium (see 6)	Johnson & Shockley[4] (thermionic emission)		(112) (011) (111) (001)		900°K for atoms on (112) and (011) 700°K for atoms on (111) and (001)	
(9)	Tungsten (see 1)	Thorium f.c.c. $a = 5.0$ Å $d = 3.6$ Å	Benjamin & Jenkins[19]	(110)	(211)	Starts at 870°K	Starts at 2100°K from all parts	
(10)	Molybdenum (see 5)	Thorium (see 9)	Benjamin & Jenkins[19]	(110)	(211)	Starts at 870°K	Starts at 1170°K except from (211)	

TABLE V *Continued*

	Metal Surface	Adsorbed Atoms	Observer	Least Emission from	Adsorbed Atoms Prefer	Migration Temperature of Adsorbed Atoms	Evaporation Temperature of Adsorbed Atoms	Work Function in Electron Volts
(11)	Tungsten (see 1)	Sodium b c c $a = 4.24$ Å $d = 3.72$ Å	Martin thermionic emission	—	100 and triangles with corners at 211	—	Starts at 1200 K Atoms on 100 leave last	
(12)	Tungsten (see 1)	Sodium see 11	Benjamin & Jenkins	110	211 100	Thick films migrate at room temperature monolayers and patches at about 500 K	Starts at 600 K Completely evaporated at 950 K	
(13)	Molybdenum (see 5)	Sodium (see 11)	Benjamin & Jenkins	110	211 100	As for tungsten (see 12)	As for tungsten (see 12)	
(14)	Tungsten see 1	Potassium b c c $a = 5.33$ Å $d = 4.46$ Å	Haefer					2.2 for very thick layers 1.56 for optimum thickness
(15)	Tungsten (see 1)	Copper f c c $a = 3.6$ Å $d = 2.54$ Å	Ashworth		Details of observations are discussed in Section IV.4			

barium are preferentially adsorbed on certain tungsten faces, as observed experimentally [see Table V (3), (5), (6), and (8)] and that such adsorption can only occur if these contaminants are atomic. Field emission microscopy can contribute useful information on adsorption processes, and some examples are cited in the next section.

3. Factors Influencing Adsorption Processes

The factors which may influence foreign atom adsorption processes on clean metals are:

(1) The specific surface energy of the particular crystal face which will influence the degree of binding between the surface and adsorbed atoms.

(2) Whether the adsorbed layer of atoms is coherent or incoherent.

(3) The nature of the binding forces between the adsorbed atoms among themselves and between adsorbed atoms and the surface atoms.

(4) The relative sizes of the adsorbate atoms and those of the adsorbent.

(5) The crystal surface structure of the adsorbent.

At a given temperature an adsorbing surface can only acquire a coherent film of foreign atoms by condensation from the vapor phase if the vapor pressure is greater than a certain value—the saturation pressure. This may be considered in another way. The adsorbed film will not be coherent if its temperature is such that its constituent atoms can acquire energy greater than the sum of the binding energies (*a*) between the adsorbed atom and its like neighbors and (*b*) between the adsorbed atom and its neighbors in the adsorbent surface. For (*b*) to be important it may be that the adsorbed atom should come into the closest "contact" with the adsorbent lattice.

It is usually assumed that adsorbent surfaces having high specific surface energies can most firmly adsorb foreign atoms. This is true, however, only if the adsorbed atoms have approximately the same cross section as the adsorbent surface atoms. If the foreign atoms have a larger cross section and the number of their adsorbent atom neighbors is important for binding then the binding energy may be greatest for adsorption on surfaces of low specific surface energies since on such surfaces the number of adsorbent atom neighbors may be large.

The first criterion for preferential adsorption is that the adsorbed foreign atom layer shall fit structurally to the adsorbent surface with the minimum of lattice strain. The greater the strain the more readily the adsorbed atoms may be evaporated from the surface. A complete analysis of the possible fits which might occur between adsorbent and adsorbate lattices is difficult, but it is relatively straightforward to determine whether or not foreign atoms are likely to form a coherent layer on a particular tungsten crystal face.

Let us consider the (001) face of the body-centered cubic tungsten with lattice parameter a = 3.16 Å and tungsten atom diameter d = 2.74 Å. Figure 5 illustrates a number of possible ways in which adsorbed atoms can form a coherent layer on the tungsten (001) face. The surface structure of the tungsten is shown in each case as a quadratic network of tungsten atoms, represented by thin line circles. The adsorbed atoms, represented by thick circles, are shown in each case to fit in repeating

TABLE VI

(a) Parameters and Diameters for Adsorption on Tungsten (001) and (011)

Integer = n (see Table VII)	0	1	2	3
Lattice parameter = a'	3.16	4.47	6.32	8.92
Atomic diameter = $d(b)$ (Body-centered cubic)	2.74	3.82	5.48	7.61
Atomic diameter = $d(f)$ (Face-centered cubic)	2.23	3.16	4.47	6.32

(b) Parameters and Diameters for Adsorption on Tungsten (111) Face

	1		2		3	
Integer = n (see Table VII)	a	b	a	b	a	b
Lattice parameter = a'	3.16	3.65	6.32	7.3	9.18	10.95
Atomic diameter = $d(b)$ (Body-centered cubic)	2.74		5.48		8.22	
Atomic diameter = $d(f)$ (Face-centered cubic)	2.23	2.57	4.47	5.14	6.72	7.71

TABLE VII

Tungsten Crystal Face	Adsorbed Lattice Structure	Adsorbed Lattice Parameter	Adsorbed Atom Diameter
(001) & (011)	b.c.c.	$a\cdot(\sqrt{2})^n$	$d\cdot(\sqrt{2})^n$
	f.c.c.	$a\cdot(\sqrt{2})^n$	$d\cdot\frac{(\sqrt{2})^{(n+1)}}{\sqrt{3}}$
(111)'	b.c.c.	$n\cdot a$	$n\cdot d$
	f.c.c.	$n\cdot a$ $\frac{2}{\sqrt{3}}n\cdot a$	$\frac{\sqrt{2}}{\sqrt{3}}\cdot n\cdot d$ $\frac{2\sqrt{2}}{3}\cdot n\cdot d$

Note. a = tungsten lattice parameter = 3.16 Å.
d = tungsten atom diameter = 2.74 Å.

array into the tungsten pattern. Figures 5*a* and *d* show two possible adsorbed coherent films with the same lattice parameter as tungsten, the former having body-centered cubic structure, the latter having a face-centered cubic structure. Other possible fitting structures are shown in Figs. 5*b*,*c*,*e*,*f*, and *g* and the corresponding lattice parameters and atomic diameters are summarized in Table VI*a*. It is seen that for each possible value of lattice parameter of the adsorbed lattice there are both body-centered cubic and face-centered cubic structures distinguished only by their atomic diameters.

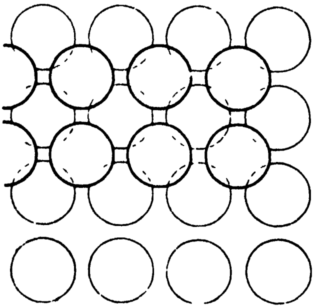


FIG. 5*a*. $a' = 3.16$ Å; $d(b) = 2.74$ Å; b.c.c. film.

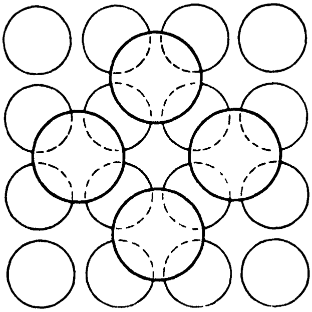


FIG. 5*b*. $a' = 4.47$ Å; $d(b) = 3.82$ Å; b.c.c. film.

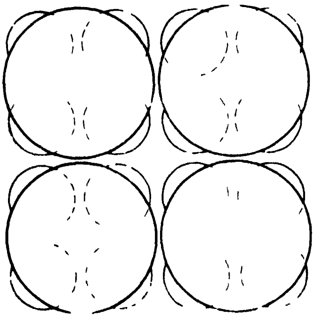


FIG. 5*c*. $a' = 6.32$ Å; $d(b) = 5.18$ Å; b.c.c. film.

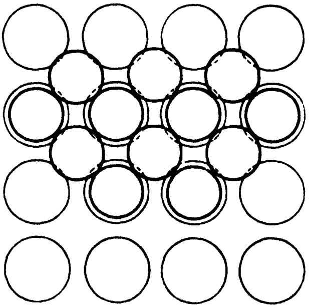


FIG. 5*d*. $a' = 3.16$ Å; $d(f) = 2.23$ Å; f.c.c. film.

FIG. 5. Adsorbed films on tungsten crystal (001) faces.

Only certain adsorbed atoms (thick-line circles) can form coherent layers on a tungsten (001) surface, shown here as a quadratic network of tungsten atoms (thin-line circles). Only discrete values of adsorbed film lattice parameter (a') and atomic diameter [$d(b)$ or $d(f)$] are allowed (see Table VI*a*) and these are functions of the tungsten lattice parameter (see Table VII). Some films have body-centered cubic structure, others are face-centered cubic.

Figure 6 illustrates three possible adsorbed layer formations on the (011) tungsten face. The adsorbed films have body-centered cubic structures in (*a*) and (*b*) and a face-centered cubic structure in (*c*). The values of lattice parameters and atomic diameters for coherent adsorption on the (011) face are given in Table VI*a*.

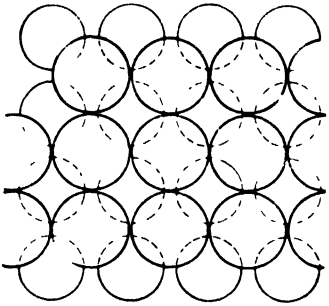


Fig. 5*e*. $a' = 4.47$ Å; $d(f) = 3.16$ Å, f.c.c. film.

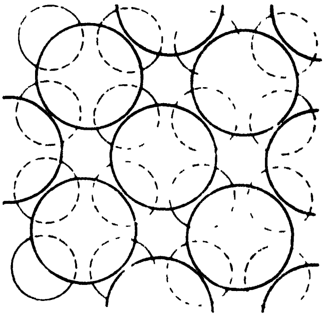


Fig. 5*f*. $a' = 6.32$ Å; $d(f) = 4.47$ Å; f.c.c. film.

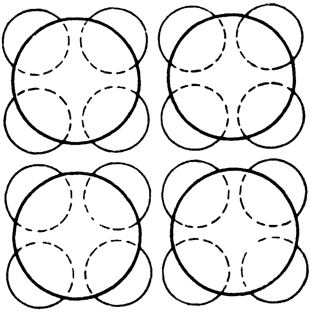


Fig. 5*g*. $a' = 8.92$ Å; $d(f) = 6.32$ Å; f.c.c. film

Fig. 5. (*Continued.*)

Figure 7 shows possible adsorbed layer formations on the (111) tungsten face, body-centered cubic structure in (*a*) and two face-centered cubic in (*b*) and (*c*). The appropriate values for parameters and diameters are given in Table VI*b*.

General expressions for the values of the crystal parameters and atomic diameters of the b.c.c. and f.c.c. foreign materials for coherent adsorption to occur on the (001), (011), and (111) faces of any body-centered cubic lattice are given in Table VII (p. 21).

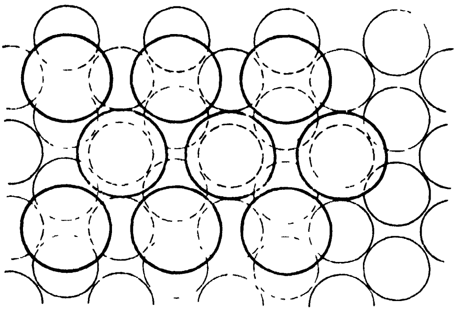


FIG. 6*a*. $a' = 4.47$ Å, $d(b) = 3.82$ Å; b.c.c. film.

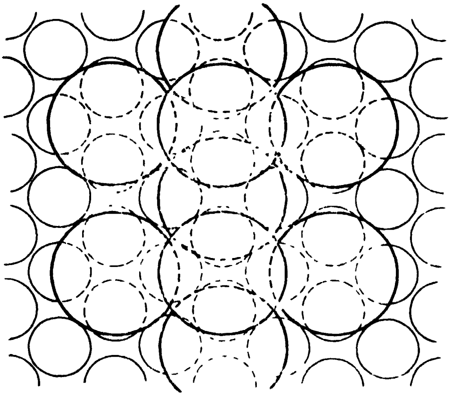


FIG. 6*b* $a' = 6.32$ Å, $d(b) = 5.48$ Å, b.c.c. film

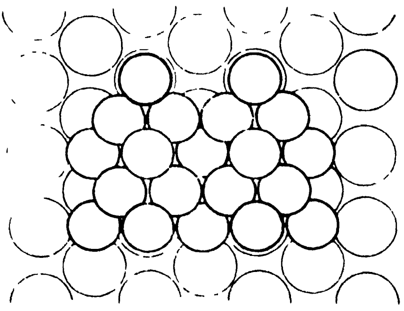


FIG. 6*c*. $a' = 3.16$ Å, $d(f) = 2.23$ Å, f.c.c. film.

FIG. 6. Adsorbed films on tungsten crystal (011) faces.

Only certain adsorbed atoms (thick-line circles) can form coherent layers on a tungsten (011) surface (thin-line circles). Only discrete values of adsorbed film lattice parameter (a') and atomic diameter [$d(b)$ and $d(f)$] are allowed (see Table VI*a*) and these are functions of the tungsten lattice parameter (see Table VII). Some films have body-centered cubic structure, others are face-centered cubic.

It is of interest to consider the lattice parameters and atomic diameters of those elements which have received attention as adsorbates in the thermionic and field emission observations listed in Table V and to seek confirmatory evidence for the information contained in Tables VI

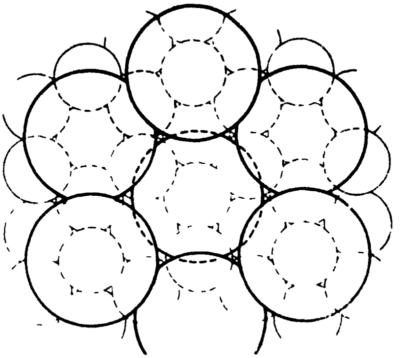


Fig. 7*a*. $a' = 6.32$ Å; $d(b) = 5.48$ Å; b.c.c. film.

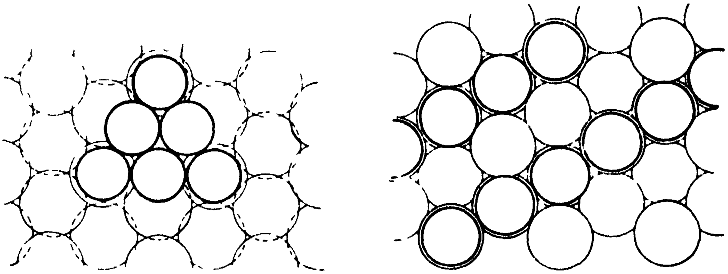


Fig. 7*b*. $a' = 3.16$ Å, $d(f) = 2.23$ Å, f.c.c. film.

Fig. 7*c* $a' = 3.65$ Å; $d(f) = 2.57$ Å; f.c.c. film.

Fig. 7. Adsorbed films on tungsten crystal (111) faces.

Only certain adsorbed atoms (thick-line circles) can form coherent layers on a tungsten (111) surface (thin-line circles). Only discrete values of adsorbed film lattice parameter (a') and atomic diameter [$d(b)$ and $d(f)$] are allowed (see Table VI*b*) and these are functions of the tungsten lattice parameter (see Table VII). Some films have body-centered cubic structure, others are face-centered cubic.

and VII. The lattice parameters and atomic diameters of tungsten and molybdenum differ by less than 1% so that they may be considered together as one type of adsorbent. The lattice parameters and atomic diameters of these adsorbates may be compared with the values required for adsorption according to Table VI and the results are summarized in Table VIII.

TABLE VIII

Foreign Atoms	Lattice Constant*	Atomic Diameter*	Possible Adsorption (from Table VI)	Observed Adsorption (from Table V)	Reference Table V
Barium (b.c.c.)	5 02	4 2	Not on (001) (011) or (111)	Not on (001) (011) or (111)	(3), (4), (5)
Cesium (b.c.c.)	6 2 (6 32)	5 4 (5 48)	Adsorption on (001); (011) & (111)	Adsorption on (001); (011) & (111)	(6), (7), (8)
Thorium (f.c.c.)	5 0	3 6	Not on (001) (011) or (111)	Not on (001) (011) or (111)	(9) & (10)
Sodium (b.c.c.)	4 24 (4 47)	3 72 (3 82)	Adsorption on (001) & (011) Not on (111)	Adsorption on (001) Not on (011) or (111)	(11), (12), (13)
Copper (f.c.c.)	3 6 (3 65)	2 54 (2 57)	Adsorption on (111) Not on (001) or (011)	Adsorption on (111) Not on (001) or (011)	(15); see also IV, 4

* Lattice constants and atomic diameters in parentheses are the values required theoretically (Tables VI and VII) for adsorption to occur.

Table VIII shows that the few experimental results so far obtained generally confirm the hypothesis. Individual cases require individual treatment. It is possible, for example, to explain the observation that sodium was adsorbed on the (001) face of tungsten, although it was not found on (011), see Table V, (12) and (13), and although the conditions for adsorption according to Tables VI and VII are identical. It must be remembered that Tables VI and VII only consider how one structure geometrically fits another. Physically this is but one contribution and other factors may be equally important in determining whether a given adsorbate is in equilibrium on an adsorbent surface under stated conditions of temperature and adsorbate vapor pressure. A second contribution is the binding energy between the adsorbent and adsorbate. The relative magnitudes of binding energies, for adsorbate atoms on the various adsorbent crystal faces, is determined by their relative numbers of adsorbent atom neighbors. A sodium atom on a tungsten (001) face (Fig. 5*b*) will experience the attractive influence of four near tungsten neighbours, while a sodium atom on a tungsten (011) face (Fig. 6*a*) will have only one or two near neighbors depending upon its precise situation. It follows that the sodium atom and the sodium lattice are less tightly bound to the (011) face than to the (001) face and we may conclude that there is a greater probability for adsorption on the latter as observed experimentally.

4. *Field Emission Studies of the Adsorption of Copper on Tungsten*

As an example of the type of field emission study which is possible with the field emission microscope, the adsorption of copper atoms on a clean tungsten surface[27] will be described and illustrated in detail. The construction of the microscope is shown in Fig. 8. The fluorescent screen covers the larger part of the inner wall of a spherical bulb in which is mounted the emitting specimen. This consists of a suitably prepared tungsten wire point spot welded to a hairpin heating filament. The bulb carries a connecting tube for its evacuation, a side tube which is immersed in liquid air and a side tube containing a source of the adsorbent material —in this case a copper-coated tungsten filament.

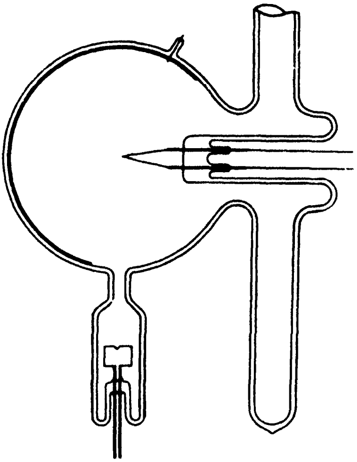


FIG. 8. The field emission microscope used in preparing Figs. 9 and 10 showing the fluorescent screen anode and a tungsten emitter mounted on the heating filament. The copper source is contained in the left-hand tube, and the tube to be immersed in liquid air is on the right.

Briefly, the procedure is as follows:

(1) Evacuate and bake out the apparatus including degassing the copper-coated tungsten and the tungsten emitting point and filament.*

(2) Immerse the side tube in liquid air.

(3) Clean the tungsten point by heating to 3000°K for a few seconds.

(4) Apply several kilovolts between anode (fluorescent screen) and emitter and observe fluorescent image for clean tungsten crystal. (If the

---

* Good vacuum technique is an essential part of this work. A general account is given by Dushman[29] and experimental details have been published by Anderson and Nottingham.[30]

image is not a symmetrical pattern it can often be improved by bombarding the emitting point with argon ions by applying an accelerating potential between point and screen with the bulb filled with argon at 10^{-5} cm pressure.)

(*a*) Emission for 8 kv applied, from a "built-up" clean tungsten surface produced by flashing at 3000°K with 3w applied between point and screen.

(*b*) Copper atoms directed from source strike the underside of the emitter and on arrival produce spots of emission on screen. The emission from the underlying tungsten surface is suppressed. 7 kv applied.

(*c*) After one hour at room temperature. 7 kv applied.

(*d*) After heating to 900°K for fifteen minutes. 7.5 kv applied.

(*e*) After heating to 1000°K for five minutes. 6.5 kv applied.

(*f*) After heating to 1000°K for three minutes. 7.5 kv applied.

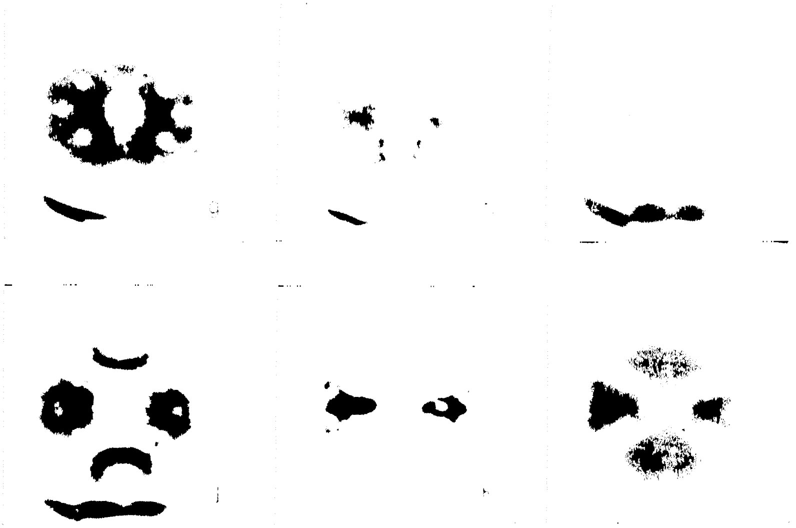
FIG. 9. The deposition of copper atoms on clean tungsten, their migration, distribution and final removal as illustrated by field emission microscopy. Diameter of emitting point approximately 10^{-4} cm. Diameter of fluorescent screen 10 cm. Vacuum better than 10^{-11} cm Hg.

(5) Deposit copper atoms from source on to the tungsten crystal surface and observe the emission image.

(6) Modify the distribution of copper over the surface by heating the tungsten. This either accelerates migration of the copper, or partially or completely evaporates it from the surface. The effects of different temperatures may be studied.

(7) By flashing the tungsten at 3000°K all the copper may be evaporated, leaving the clean tungsten surface.

The results of such an experiment are illustrated in Fig. 9. The first photograph (*a*) shows the fluorescent image of the clean tungsten surface* whose (011) axis is normal to the central nonemitting region which is the corresponding (011) face. Copper atoms were directed from the source to strike the underside of the emitter and (*b*) shows emission from the copper atoms deposited on the lower half of the microcrystal with the underlying tungsten surface obscured. Although the tungsten surface was at room temperature, the scintillating emission distribution shows



(*g*) After heating to 900°K for seven minutes Probably partial monatomic layer of copper. 8 kv applied.

(*h*) Same conditions as (*g*) but only 6 kv applied.

(*i*) After heating to 1200°K for ten seconds. 8 kv applied.

(*j*) After heating to 1200°K for ten seconds. 8 kv applied.

(*k*) After heating to 1200°K for five minutes. Only traces of copper left on (111) and (122) faces. 8 kv applied.

(*l*) After flash at 3000° K with kv applied, leaving "built-up" clean tungsten pattern. 8 kv applied.

FIG. 9. (*Continued.*)

---

* Actually the image shown is that for tungsten cleaned by flashing at about 3000°K with a potential field applied at the surface. While at this temperature adsorbed impurities volatilize; the field causing migration of surface tungsten atoms themselves. This migration accentuates surface edges, sharpening the definition of crystal face boundaries. This type of image has been termed the "built-up" pattern by Benjamin and Jenkins.[18,19]

the adsorbed copper atoms to be in a state of energetic migration, settling from a randomly deposited cluster of atoms to a more stable structure. Photographs (*b*) and (*c*) illustrate stages in the room temperature migration process. This migration and settling process could be accelerated by raising the surface temperature. The rate of fluctuation of scintillation of the image was taken as an indication of migration rate and was appreciable even at room temperature.

The redistribution process was accelerated by heating the specimen to 900°K, and the change is illustrated in (*d*). At a later stage during heating, all the copper crystallites except two or three in the vicinity of the (100) boundary were lost by evaporation (*e*). The reappearance of the symmetrical image associated with the underlying tungsten structure was observed in the background and became more apparent after further heating (*f*).

The equilibrium condition at 900°K is shown in (*g*). The emission arises principally from a firmly bound film of copper adsorbed preferentially on certain tungsten crystal faces. This film could only be disturbed by raising the temperature to 1200°K when it was evaporated progressively [(*h*),(*i*),(*j*)] from the (113), (116), and (013) faces. The copper on the (111) and (122) faces remained at 1200°K (*k*) and was only evaporated at about 3000°K (*l*) when the original clean tungsten emission image reappeared.

We have already considered in Section IV, 3 the necessary conditions for coherent adsorption in terms of structural fits between the adsorbent surface and the adsorbate lattice, and in the case of copper on tungsten it was found that on these terms strong adsorption should be possible on the tungsten (111) faces but not on the (001) and (011) faces The experimental observations described above are in agreement with these conclusions. Further work is required along these lines to establish whether similar agreements exist for adsorption or nonadsorption on the more complex faces.

5. *Field Emission Studies of More Complex Surface Processes*

We have described how the field emission microscope may be used to study the adsorption processes in which atoms of one element condense on crystal faces of another element. More complex processes such as the slow oxidation of a thin metallic film adsorbed on a tungsten surface may be investigated by the same technique, but unfortunately the interpretation of the results is extremely difficult. The problem reduces to the analysis of a two-dimensional observation to describe a three-dimensional process and has no precise solution.

As an example of this, the results of studying the slow oxidation of a copper film adsorbed on a clean tungsten surface[27] are illustrated in Fig. 10. The field emission microscope described in Section IV, 4 was used, and the rate of oxidation of the copper film was retarded by working in a vacuum of 10^{-10} cm Hg (see Section V). A technique for preparing the clean copper film was developed from the experience gained in the work described in Section IV, 4. The resultant film was confined to particular faces of the tungsten surface and only a few atoms thick. As was shown previously (Fig. 9), a film thicker than this does not retain the structure of the intimately adsorbed layers but tends to grow crystallites resulting in complex emission images. The limited thickness of the film and the fact that it is localized on certain tungsten faces are limitations of the technique.

A tungsten point specimen was prepared with its (011) crystallographic direction in the axis of the wire and its surface was completely outgassed by the process described in Section IV, 4. The pressure in the system which had been previously flooded with oxygen, was 10^{-10} cm Hg, which meant that in ten minutes a cleaned metal surface would become seriously contaminated with adsorbed oxygen (see Section V). A technique was developed for preparing the copper film on cleaned tungsten in less than one minute in order to avoid serious oxidation. The copper was then slowly oxidized by adsorption at room temperature. After an equilibrium condition had been established, the temperature of the specimen was raised stage by stage, and the resultant changes in the emission images were recorded on 35-mm film from which the photographs in Fig. 10 are taken.

The emission image from the cleaned "built-up" tungsten surface (*a*) closely resembled that analyzed in Section III. By maintaining the temperature of the cleaned tungsten surface at about 1000°K during the deposition of the copper film, the required adsorbed layer could be quickly prepared as shown in (*b*). The progressive changes in the image as the copper adsorbed oxygen are illustrated in (*c*) to (*f*). The equilibrium condition for the oxygen-copper structure at room temperature was reached after about twelve minutes (*g*). No further modification occurred during the next three minutes. In this condition the emission occurs principally from the (012) face and the face boundaries and edges of (122), (233), and (111). The (023), (116), (013), (113), (001), (112), and (011) faces do not emit. The presence of an oxygen-rich structure on these faces may account for the reduced emission.

By heating the specimen first to 700°K and then to 900°K for ten seconds each [(*h*) and (*i*)] the relative emission intensities from the regions surrounding the (111) face and from the (111) face itself were enhanced. The initial concentration of copper on these surfaces was *high, as shown in (b), and it would appear that the enhanced emission was from the copper atoms which have migrated to the surface of the copper-oxygen lattice through the oxygen-rich structure.*

Further heating at 900°K resulted in some loss of copper from the (111) face and the surrounding regions and a migration of copper to the surface over the (100) face and surrounding regions (*j*). The "built-up" image of (*k*) confirms this. Further observations suggest that the copper migrates to the surface at various places at different temperatures, although no details of the processes can be deduced on account of the complexity of the surface structure both in content and geometry. After heating to 2500°K, however, emission is still intense from the (111) face and surrounding regions (*k*), which suggests the presence of copper on

(*a*) Emission, for 8.5 kv applied, from a "built-up" clean tungsten surface prepared by flashing at 3000°K with 3 kv applied between point and screen.

(*b*) Copper film deposited from vapor phase while tungsten surface was maintained at 1000°K. 8 kv

(*c*) After exposure to oxygen at 10^{-10} cm Hg pressure for two minutes. Oxygen has been adsorbed by the copper film. 10 kv.

(*d*) Emission modified after further two minutes 10 kv

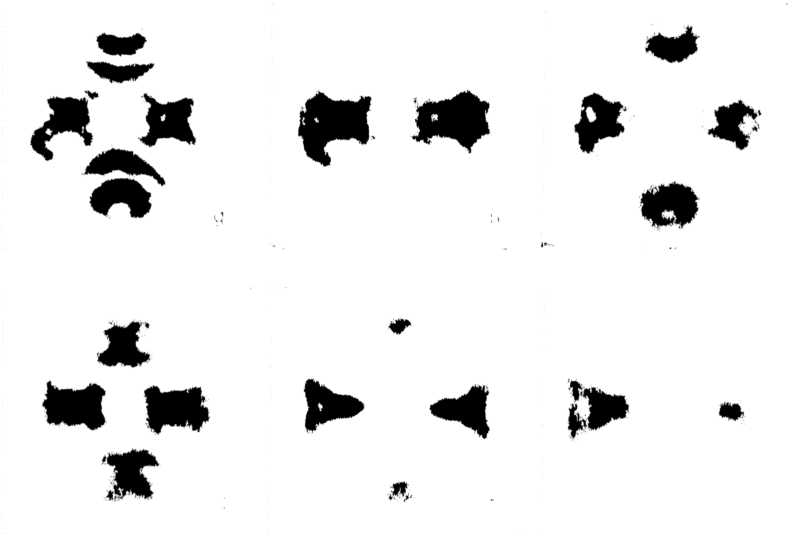
(*e*) After further two minutes 10 kv.

(*f*) After further two minutes. 9.5 kv.

FIG. 10. A complex adsorption process. The oxidation of a copper film on a clean tungsten substrate Diameter of emitting point approximately 10^{-4} cm Diameter of fluorescent screen 10 cm Vacuum 10^{-10} cm Hg.

these faces. This observation is confirmed by the results illustrated in Fig. 9*k*, which shows that some copper is firmly bound to these faces even at 1200°K and possibly higher. Heating to 3000°K removes both copper and oxygen, leaving the clean tungsten (*l*).

The observations described above show in a qualitative way the changes in the emission detail due to changes in the structure and content of the oxygen-copper film formed on the tungsten substrate. The possibilities of this type of field emission study are obviously seriously limited by its failure to provide information relating to the third dimension.



(*g*) After further two minutes 9.5 kv.
(*h*) After heating to 700°K for ten seconds Emission enhanced around (111) faces 9.5 kv.
(*i*) After heating to 900°K for ten seconds 9.5 kv.
(*j*) After heating to 900°K for ten seconds 9.5 kv.
(*k*) After flashing at 2000°K 9 kv
(*l*) After flashing at 3000°K with 3 kv applied "Built-up" clean tungsten image 9 kv

FIG. 10. (*Continued.*)

V THE FIELD EMISSION MICROSCOPE AS A HIGH-VACUUM GAGE

Electron emission images of clean metallic surfaces have been discussed in Section III, and in particular the image from a tungsten surface was described. Since the atomic structure of the surface of the specimen cannot be determined in any detail before it is introduced into the microscope system, on account of its size, the type of emission image obtained is not predictable. Different images observed vary from wholly unintelligible distributions of bright and dark patches to perfect patterns whose symmetry may be directly related to the body-centered cubic structure for the tungsten lattice.

When the surface of a tungsten microcrystalline specimen is thoroughly cleansed of adsorbed foreign atoms by heating it to 3000°K in a vacuum better than 10^{-8} mm Hg, the intensity distribution of the emission image is found to change over a period of some seconds or minutes. The original sharp image from the clean surface becomes blurred, and the intensity distribution is modified by the random appearance of small relatively brightly emitting spots, while the general emission intensity slowly diminishes. It is necessary to increase the anode potential to maintain the original intensity. The phenomenon may be repeated at will and is in fact the simplest type of field emission microscope observation.

If a known gas pressure in the range 10^{-6} to 10^{-10} mm Hg, is introduced into the microscope and the time is taken for the image to be completely modified to a new equilibrium pattern noted, then a pressure ten times greater will be found to take one-tenth of the time to reach the same equilibrium condition. Such observations are in accord with the predictions of the kinetic theory of gases. Gas atoms or molecules will be readily adsorbed by a previously cleaned metal surface. The rate at which this contamination occurs will be determined by the rate at which the gas particles strike the surface in their random flight which in terms of the kinetic theory is proportional to the gas pressure. Table IX is derived for oxygen and shows for a range of gas pressures the number of gas particles striking each square centimeter of a surface every second, the number striking a typical target area of 10^{-5} cm square every second, and the time taken for the target to become completely covered with adsorbed particles, assuming that each one striking the surface finds a site available for adsorption and sticks. Actually some atoms or molecules will collide elastically with the surface and escape adsorption, and a correction factor is required. This correction will depend on the nature of the gas particles and of the adsorbent surfaces and a number of other factors. It need not concern us here since only relative orders of magnitude are of interest.

The field emission image from an initially clean tungsten surface, for example, is modified by the adsorption of gas particles, and the time taken for a particular degree of modification or surface contamination

TABLE IX

Gas Pressure mm Hg	Number of Impacts per sq cm per sec	Number of Impacts on Typical Target (10^{-10} sq cm) per sec	Time Taken for Gas Particles to Cover Completely the Target Area
10^{-5}	4×10^{15}	4×10^{5}	0.06 second
10^{-6}	4×10^{14}	4×10^{4}	0.6 second
10^{-7}	4×10^{13}	4×10^{3}	6 seconds
10^{-8}	4×10^{12}	4×10^{2}	1 minute
10^{-9}	4×10^{11}	40	10 minutes
10^{-10}	4×10^{10}	4	100 minutes
10^{-11}	4×10^{9}	< 1	1000 minutes

to occur is inversely proportional to the residual gas pressure. It follows that the field emission microscope can itself be used as a sensitive high-vacuum gage with a useful range from 10^{-6} mm to 10^{-11} mm Hg. In spite of the immediately apparent possibilities, its only practical use is found in the field emission observations themselves.

It is worth noting in passing that most surface emission and adsorption experiments today rely for their high-vacuum measurements on this type of phenomenological observation. In measuring work functions of clean metals, for example, it is essential to work *in vacuo* better than 10^{-10} mm Hg. This may be checked by observing the changing work function of the surface after its initial cleaning until an equilibrium value is reached. The time taken gives an order of the degree of high vacuum obtaining.

VI. THE RESOLVING POWER OF THE FIELD EMISSION MICROSCOPE

1. *Observations on Individual Gas Atoms and Molecules*

Occasionally an etched specimen produces an emission image which consists of large dark areas with a few brightly emitting regions randomly distributed around them. Such an image may arise from an extremely small specimen having relatively large flat faces surrounded randomly by sharp edges and corners. In the same way as that already described for the observations on symmetrical patterns in Section V, the rate of random arrival of relatively bright spots on the dark areas of the image are found to be pressure dependent. Frequently with this type of image the spots are much larger than those scintillations discussed in Section V arising from greater magnification, and their behavior can be followed in greater detail.[27] Single spots appear initially after the surface has been cleaned, but after a fraction of a second or more, they split into two halves which remain close together, usually rotating about each other for a similar period before either disappearing completely or splitting into two quite separate spots which move off independently and at random over the image. Ultimately as the surface becomes more and more contaminated, these processes become less obvious.

These observations can be explained by assuming that the spots which appear initially are oxygen molecules which are subsequently dissociated into their constituent atoms. These atoms are loosely adsorbed in adjacent sites on the tungsten lattice, and at room temperature there is considerable probability that interchange of positions can occur, resulting in an apparent rotation of the two spots about each other. The two individual atoms can become completely dissociated and move randomly and rapidly over the non-emitting surface to be lost either in the bright emission from an edge or corner or by evaporation from the surface. This motion corresponds to the mobility of oxygen atoms across close-packed tungsten lattice faces at room temperature. This behavior is recognized as similar to that which has been described by Roberts and others[25] as a deduction from measurements of the adsorption of oxygen molecules on tungsten surfaces.

The evidence in the above observations is inconclusive but further experiments may be carried out in which, during baking, the whole vacuum system is flooded with gases other than oxygen. Hydrogen and argon have been used.[27] The degree of vacuum obtained before and after the flooding process should be of the order of 10^{-10} mm Hg. The results with hydrogen are found to be similar to those already described for oxygen, but for argon no splitting of the bright spots is observed. This provides further evidence that the splitting which only occurs for the diatomic molecules is the dissociation of the molecule into its constituent atoms.

These phenomena can be observed only when they occur on atomically flat crystal faces which do not themselves emit electrons when subjected to the normal potential fields used in this work (several kilovolts between crystal and screen). This suggests that when a single atom or molecule lands on this type of surface, the smooth contour and its associated potential field are disturbed, and electrons can be emitted preferentially from the region surrounding the disturbance.

The photograph in Fig. 11 is a record of a typical observation showing the emission from the neighborhood of a partially dissociated oxygen molecule momentarily occupying two adjacent sites on a tungsten crystal face. The photograph is one of a series of 300 random exposures of 0.5 second duration using 35-mm film. This procedure was necessary on account of the continuous appearance, movement, and disappearance of the image spots. The other emission spots in the photograph are due to other atoms or undissociated molecules, while the larger bright patches are due to emission from edges and corners on the otherwise flat surface.

In order to distinguish between the two oxygen atoms occupying adjacent sites on the tungsten face, the resolving power of the microscope must be of the order of the separation of nearest neighbor tungsten atoms in the lattice, 2.74 Å. The next section considers whether this is theoretically feasible.

FIG. 11 Photograph of the image produced by the field emission from the vicinity of an oxygen molecule (center) adsorbed on a tungsten surface

2. Theoretical Aspects of Resolving Power

In attempting to calculate the resolving power of the field emission microscope, the problem may be approached from two angles. The emitted electron beam may be treated either as a stream of particles obeying mechanical laws or as waves radiating from the source analogous to light waves emitted from and diffracted at an aperture. We shall confine the discussion to the case in which two atoms are adsorbeb in adjacent sites on an otherwise smooth tungsten surface, their centers 2.74 Å apart.

a. Particle Electrons. If, in Fig. 12, r and R are the radii of curvature of the specimen and the fluorescent screen respectively, V the potential between them, V_t the electron surface kinetic energy in volts, and u_t and u_R the electron velocity components tangential and normal to the emitting surface, then, from particle mechanics,

$$\tfrac{1}{2}m \cdot u_R^2 = e \cdot V \quad \text{and} \quad \tfrac{1}{2}m \cdot u_t^2 = e \cdot V_t \tag{5}$$

The radius of the image circle of confusion is determined by the motion of the electrons in the direction normal to the common radius during their flight from emitter to screen. The electrons take R/u_R seconds to reach the screen, and the resultant radius of confusion will be $u_t \cdot R/u_R$ centimeters, giving a fluorescent spot diameter from eq. 5 of

$$D = 2R\sqrt{V_t/V} \tag{6}$$

and a resolving power of

$$d = D/M \tag{7}$$

where M is the magnification of the system.

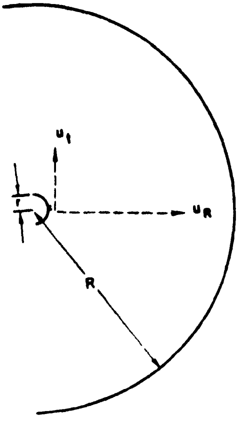


Fig. 12

In Fig. 13 are plotted values of D for V_t between 0.01 and 1000 ev and for V between 10 and 50,000 ev for a bulb radius of 5 cm. Corresponding values of d are included for magnifications from 10^5 to 10^8. To determine the appropriate values of D and d for the case of two oxygen atoms occupying adjacent sites on a tungsten face, it is necessary to determine the most likely value of V_t, the electron surface tangential energy. This is a function of the thermal energy and that resulting from any distortion of the potential field in the neighborhood of the surface disturbance (an adsorbed atom or molecule). The thermal contribution at room temperature is $kT = 0.025$ ev. The field distortion contribution according to Richter[26] is

$$e \cdot r' \cdot E \cdot \sin^2 \theta \tag{8}$$

where r' = the radius of the surface disturbance,
E = the field intensity in the vicinity of this disturbance,
θ = the angle between the normal to the surface and the direction of the distorted field in the vicinity of the disturbance.

To determine the order of magnitude of this field contribution we must consider a specific case. If emission occurs from the vicinity of a single atom or molecule of say, 2 Å effective radius, then neglecting disturbing effects it will produce an undistorted image of radius 2 mm for a system with magnification 10^7. With this magnification rectilinear emission from a tungsten crystal specimen of 5×10^{-7} cm radius or 40 tungsten atoms diameter, will completely fill the screen of radius 5 cm. A potential of 20 kv between specimen and screen produces a field intensity at the surface of 4×10^{10} volts per centimeter. From normal experience a field of 10^{10} volts per centimeter would appear to be somewhat disruptive.

It is possible that space charge effects near the emitting surface could modify this value to say 10^8 volts per centimeter in the case under consideration. We may assume a value of 15° for θ as a possible angular distortion of the field direction in the neighborhood of the adsorbed atom. Substituting these values in expression (8) the contribution to the electron energy normal to the common radius due to the surface field distortion is found to be of the order of 10 ev. Compared with this, the thermal contribution may be neglected. Substituting $V_t = 10$ ev; $V = 2 \times 10^4$, and $R = 5$ cm, into expression (6) we obtain $D = 2.2$ mm (see also Fig. 13); and from expression (7), a resolving power of 2.2 Å.

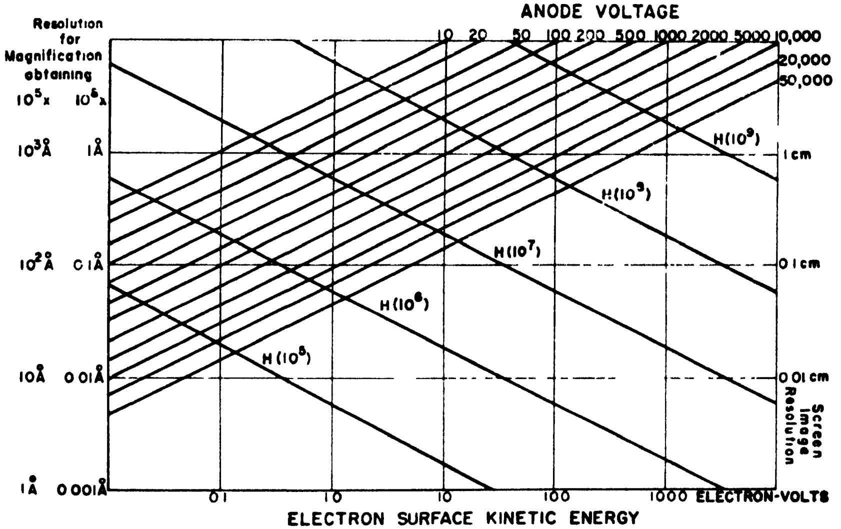


FIG. 13. The resolving power of the field emission microscope in terms of the anode voltage, the geometrical magnification of the system and the electron surface tangential energy. Lines on the graph labeled H (10^7), etc., are the Heisenberg limiting values of resolving power for the applicable magnification indicated in parentheses.

b. Wave Electrons. Benjamin and Jenkins[15] have calculated the resolving power of their emission microscope in terms of the diffraction of the electron waves as they emerge from an imaginary surface aperture, the diameter of this aperture being the resolving power. By direct analogy with the optical case, the diameter of the spot image due to diffraction is given by $1.22\lambda R/d$ where λ is the de Broglie wavelength of the electron, R is the radius of the screen, and d is the diameter of the aperture.

The ratio of intensities of successive diffraction rings is of the order of 20:1, and it is feasible to consider only the central bright diffuse spot since the circles around it are not easily detected. By combining the expressions for linear magnification and electron diffraction, an expression for the total image spot size is obtained.

$$D = R[(\lambda/d) + (d/r)] \tag{9}$$

for which D is a minimum when $d^2 = \lambda r$.

This may be applied to the specific case of two oxygen atoms occupying adjacent sites on the tungsten lattice, when the applied potential is 20 kv, corresponding to a de Broglie wavelength of 10^{-9} cm and the radius of the specimen is 5×10^{-7} cm. The minimum value of d and the resolving power is 2.2 Å, and the diameter of the image spot is 4.4 mm.

c. The Heisenberg Limiting Resolution. According to the Heisenberg uncertainty principle, the limiting resolving power of a system is equivalent to half the de Broglie electron wavelength. As already shown in Section VI, *2a*, the electron lateral energy is of the order of 10 ev, corresponding to an electron wavelength of about 4 Å. The Heisenberg limiting resolution is therefore 2 Å. Heisenberg limits are plotted for other conditions in Fig 13.

3. *Conclusions*

It has been shown in Sections VI, *2a*, VI, *2b*, and VI, *2c* that theoretically the resolving power of the system discussed in Section VI, 1 may be as good as 2.2 Å. As stated earlier, in order to distinguish between the emission images from the vicinities of two oxygen atoms adsorbed on tungsten, the resolving power would have to be better than 2.74 Å. This has been shown to be possible by using a sufficiently small emitting surface and a high accelerating potential. It should be noted that only twice in the course of a year's experiments were conditions favorable for the observation of images of the type illustrated in Fig. 11, and since the preparation of specimen surfaces as small as 5×10^{-7} cm radius is purely a matter of chance, the method cannot readily be made available for general use.

From the practical point of view cathode points of radius 10^{-5} cm are more easily prepared and although the corresponding resolving power is not so good, Müller has recently shown[31] that it is good enough to enable the general structure of fairly large molecules to be studied. With such a point he has obtained a magnification of nearly 3×10^6 and a resolving power of 7.7 Å for images of copper phthalocyanine *molecules which resemble in appearance the structure chemically assigned to them (Fig. 14).*

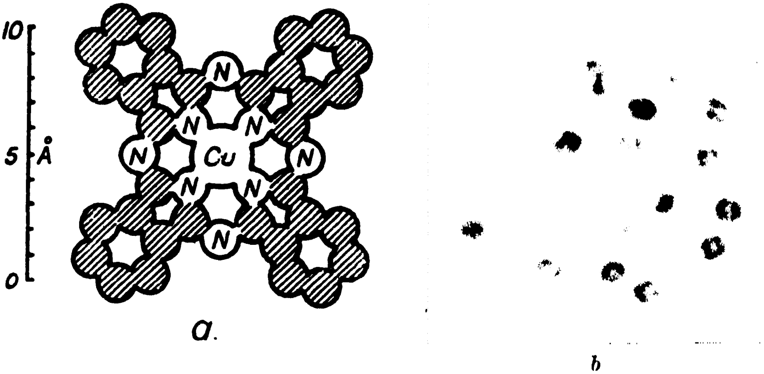


FIG. 14. (*a*) The structure of the copper phthalocyanine molecule (according to R. P. Linstead) and (*b*) the field emission images of a number of these molecules on a tungsten surface, showing the quadrupartite detail. (By permission of Dr. E. W. Müller.[13])

ACKNOWLEDGMENTS

Some of the results described in Sections IV, V, and VI, were obtained by the author at the University of Bristol while on leave of absence from Metropolitan-Vickers (1945–7). In this connection he is greatly indebted to Dr. J. W. Mitchell who proposed the work, and to Dr. R. W. Sillars for his encouragement. He appreciates also the assistance afforded by his wife in the preparation of illustrations for this review.

Finally, the author wishes to thank Sir Arthur P. M. Fleming, C.B.E., D.Eng., Director and Mr. B. G. Churcher, M.Sc., M.I.E.E., Manager of the Research Department, Metropolitan-Vickers Electrical Co., Ltd., for permission to publish this article.

REFERENCES

1. *Thermionic Experiments:*
   Brüche, E., and Johannsen, H. *Z. Physik*, **84**, 56 (1933).
   Johnson, R. P., and Shockley, W. *Phys. Rev.*, **49**, 436 (1936).
   Yerzley, F. L. *Phys. Rev.*, **50**, 610 (1936).
   Johnson, R. P. *J. Applied Phys.*, **9**, 508 (1938).
   Martin, S. T. *Phys. Rev.*, **56**, 947 (1939).
   Nichols, M. H. *Phys. Rev.*, **57**, 297 (1940).
2. *Photoelectric Experiments:*
   Pohl, J. *Z. tech. Physik*, **15**, 579 (1934).
   Mahl, H., and Pohl, J. *Z. tech. Physik*, **16**, 219 (1935).
   Brüche, E. *Z. Physik*, **98**, 77 (1935).
3. Herring, C., and Nichols, M. H. *Revs. Mod. Phys.*, **21**, 185 (1949)
4. Johnson, R. P., and Shockley, W. *Phys. Rev.*, **49**, 436 (1936).
5. Schmidt, R. W. *Z. Physik*, **120**, 69 (1942).

6. Brüche, E., and Mahl, H. *Z. tech. Physik*, **16,** 623 (1935); **17,** 81 (1936).
7. Ahearn, A. J., and Becker, J. A. *Phys. Rev.*, **49,** 879 (1936).
8. Nichols, M. H. *Phys. Rev.*, **57,** 297 (1940).
9. Robinson, C. S. *J. Applied Phys.*, **13,** 647 (1942).
10. Johnson, R. P., White, A. B., and Nelson, R. B. *Rev. Sci. Instruments*, **9,** 253 (1938).
11. Müller, E. W. *Physik. Z.*, **37,** 838 (1936); *Z. tech. Physik*, **17,** 412 (1936); *Z. Physik*, **102,** 734 (1936); **106,** 132, 541 (1937).
12. Müller, E. W. *Z. Physik*, **108,** 668 (1938).
13. Martin, S. T. *Phys. Rev.*, **56,** 947 (1939).
14. Müller, E. W. *Naturwissenschaften*, **27,** 820 (1939); *Z. Physik*, **40,** 261 (1943).
15. Benjamin, M., and Jenkins, R. O. *Proc. Roy. Soc.*, **A176,** 262 (1940).
16. Manning, M. F., and Chodorow, M. I. *Phys. Rev.*, **56,** 787 (1939).
17. Haefer, R. H. *Z. Physik*, **116,** 604 (1940); *Z. Krist.*, **104,** 1 (1942).
18. Nordheim, L. W. *Proc. Roy. Soc.* (*London*), **A121,** 626 (1928).
19. Benjamin, M., and Jenkins, R. O. *Proc Roy. Soc.* (*London*), **A180,** 225 (1942).
20. Mendenhall, C. E., and DeVoe, C. F. *Phys. Rev.*, **51,** 346 (1937).
21. Smoluchowski, R. *Phys. Rev.*, **60,** 661 (1941).
22. Wigner, E., and Bardeen, J. *Phys. Rev.*, **48,** 84 (1935).
23. Stranski, I. N., and Suhrmann, R. *F.I.A.T. Report* No. 1030 (13.1.47).
24. de Boer, J. H. Electron Emission and Adsorption Phenomena. Cambridge, 1935.
25. Roberts, J. K. Some Problems in Adsorption. Cambridge, 1939.
    Miller, A. R. Adsorption of Gases on Solids. Cambridge, 1949.
26. Stranski, I. N., and Suhrmann, R. *F.I.A.T. Report* No. 1031 (14/2/47).
27. Ashworth, F. Ph. D. thesis, University of Bristol, 1948.
28. Richter, G. *Z. Physik*, **119,** 406 (1942).
29. Dushman, S. Scientific Foundations of Vacuum Technique, John Wiley & Sons, 1949.
30. Anderson, P. A. *Rev. Sci. Instruments*, **8,** 493 (1937); *Phys. Rev.*, **54,** 753 (1938); *Phys. Rev.*, **57,** 122 (1940).
    Nottingham, W. B. *J. Applied Phys.*, **8,** 762 (1937); *Phys. Rev.*, **55,** 203 (1939).
31. Müller, E. W. *Naturwissenschaften*, **37,** 333 (1950).

Velocity Modulated Tubes

R. R. WARNECKE,* M. CHODOROW,† P. R. GUÉNARD,* AND
E. L. GINZTON†

CONTENTS

		Page
I.	Introduction	43
II.	The Basic Forms of the Klystron	45
III.	Theory of the Klystron	50
	1. Introduction	50
	2. Focusing and Beam Formation	52
	3. Theory of Gap Interaction	53
	4. Space Charge Effects	58
	5. Theory of Large Signal Effects	62
	6. Theory of Complex Bunching Systems	64
IV.	Klystron Amplifiers	66
	1. Low-Noise Amplifiers	67
	2. Low-Power Amplifiers	67
	3. Power Amplifiers	69
V.	Reflex Klystrons	71
VI.	Summary	78
	References	81

I. INTRODUCTION

Velocity modulation tubes, now known as klystrons, were invented about ten years ago. Because of the war and subsequent important peacetime applications, much intensive research and development has taken place. The various recent improvements in design and performance have been due to the advanced stage of the theory and to the understanding of the many related electronic phenomena. While some novel variations of the basic forms of klystrons can still be expected, these will no doubt fall into the class of "design improvements." It is not very likely that any more basic inventions can be made. Much of the basic research has now been published by the various groups of investigators. It is felt that it is now possible to assess the general usefulness of these devices, to review and discuss the various design features, and to present a fairly comprehensive theory which describes the various fundamental processes.

It is the purpose of this chapter to fulfil the above need and to emphasize the results which are important, especially those which may not be

---

* Laboratories de Recherches de la Compagnie Générale de T.S.F., Paris, France.
† Stanford University, Stanford, California.

well known. The authors have attempted to draw upon all information available to them; however, since the literature in this field is already abundant, the reader is often referred to articles and books of general character rather than to the original papers; this has been done on the assumption that the reader will not be a specialist. An interested reader will find greater details (in most cases) presented in the textbooks listed in the References.[1,2,3,4,4a]

Velocity modulation tubes were invented as a result of the need to exploit for various practical purposes that part of the radio spectrum situated between ½ and 100 cm in wavelength. In order to make this part of the spectrum really useful, one had to invent and develop all the usual types of devices: high-power transmitters, including oscillators and amplifiers; receiving-type amplifiers; local oscillators for superheterodyne service; frequency multipliers; electron-coupled oscillators; tubes which could be both easily modulated and tubes whose frequency stability was satisfactory; and many other special devices. It so happens that the velocity modulation principle made it practical to solve these various requirements quite well. Although some of these can also be met with other types of tubes, in many cases the klystron is either the only available device or the most convenient one.

The development of the velocity modulation tubes was not brought about by any discoveries of revolutionary principles. It was based on the inevitable conclusions from the systematic study of the electron inertia effects which were readily apparent in the existing tubes. The rapid development of resonant cavity structures, which were almost immediately incorporated into the klystron tubes, also was a gradual evolution of the usual circuits.

The practical development of klystrons was greatly accelerated by a series of papers published in a brief interval of time. The description of cavity resonators by Hansen (1939)[5,6,7] of the klystron structures combining Hansen's resonators with the principle of velocity modulation by Varian and Varian (1939),[8] and the basic theory by Webster (1939),[9,10] Hahn and Metcalf (1939)[11] and Hahn (1939)[12] were the main disclosures which were followed by a great activity during the war and detailed developments, both in theory and practice.

Similar devices were developed in several countries isolated by the war. Subsequent papers described physical structures which were nearly identical. That this should be so in spite of the wartime isolation is not too surprising; the papers of Hansen, the Varians, *et al.* were sufficiently complete and revealing that further developments followed automatically in spite of lack of further exchange of information between the various laboratories.

During the war and after, much work has been done on development of the theory. At the present time, there are not many specific subjects that require further study. The status of the theory will be described below in Section III.

II. THE BASIC FORMS OF THE KLYSTRON

The simplest form of the klystron is the two-cavity amplifier, first described in its complete form by Varian and Varian (1939). It is shown in simplified form in Fig. 1. The cathode acts as a source of electrons: between the cathode and the first cavity there is a d-c accelerating field, which, with the aid of a suitable focusing system, produces a beam of electrons which passes through two (or more) resonant cavities.

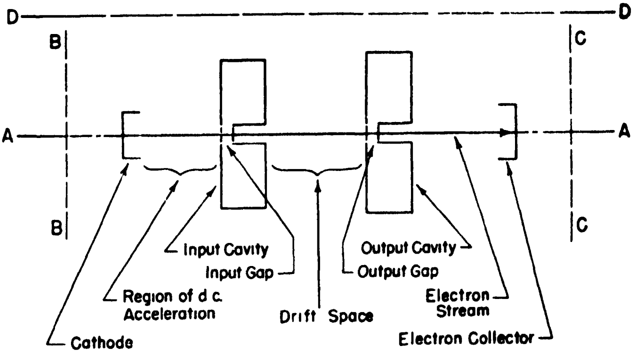


FIG. 1. Schematic view of the klystron, showing the basic parts.

Across the input (buncher) gap, a time-varying electric field is produced by introducing some radio-frequency energy into the resonant cavity (by means not shown in Fig. 1). This time-varying potential is normally small compared with the cathode-anode voltage, and the changes in velocity of electrons are therefore small. At this gap there is usually little evidence of density modulation.

In the next region, commonly called the drift or bunching space, there are no time-varying applied fields, although there may be d-c fields, either accelerating or retarding. Because of the non-uniform velocity of electrons in this space, the well-known bunching action takes place and results in grouping of electrons, thus producing density modulation at the output gap. The resultant radio-frequency current passing through the output gap is rich in harmonics, as shown in Table I. The output cavity can usually be made to have a large Q and can select any desired harmonic of the buncher frequency.

TABLE I. Maximum amplitude of *n*th harmonic of the fundamental that can be obtained at the output gap—values of current in per cent of fundamental component

Harmonic	Per Cent Fundamental Harmonic
1	100
2	83
5	64
10	52
20	42

The usual construction of klystrons corresponds to a figure of revolution about axis *A*-*A* in Fig. 1. This results in a simple structure, easily designed and constructed. Other structures are possible as, for example, the cavities can be built in the form of reentrant waveguides by extending Fig. 1 in the direction perpendicular to the plane of the paper. Other geometries are possible by generating figures of revolution about axes such as *B*-*B*, *C*-*C* or *D*-*D*. These modifications make it possible to increase the area of the cathode and other parts and to increase the power handling capacity of such systems in comparison with the simplest structure. Up to the present, very little has been made of any structure but the simplest, i.e., figure of revolution about axis *A*-*A*.

Depending upon the specific desired result, the number and type of cavities will differ. Some of these structures are shown schematically in Fig. 2 and are described below.

a. Two Cavity Amplifier. This is the simplest of all klystrons, from the viewpoint of theory and design. The efficiency of a power amplifier of this type is about 30%, and power gain is approximately 10 db. Because it is more complex than the reflex tube (see *f* below), and because its power gain and efficiency are not as good as the three-cavity tube, it is not so useful as other members of the klystron family.

b. Cascade Amplifier. A klystron with three (or more) cavities is called a cascade amplifier for obvious reasons. A cascade amplifier, when properly designed and adjusted, can have much higher power gain than that of the two-cavity klystron. Power gain of 30–40 db are common.

The cascade bunching is somewhat different from the process occurring in the simple klystron. The combined effect of the first two cavities can make the velocity grouping better. While the elementary (Webster) theory predicts an efficiency of 58%, the two-buncher system increases this up to 80%. While the absolute value of these numbers is not significant (see Section III, 6), the relative improvement is found to be correct.

If the cascade amplifier is adjusted so that it operates as a high-efficiency device, its amplification will be only moderately better than the two-cavity klystron. If operated at high amplification, its efficiency will approximate that of the two-cavity klystron.

A few unsuccessful efforts have been made to construct cascade amplifiers for preamplifier service in receivers. Although an amplification of 60 db can be obtained easily, noise figures lower than 25 db (above ideal) have not been reached. With constantly improving knowledge of noise mechanisms in high-frequency structures, it is believed that appreciably better performance can now be obtained (see Section IV, 1).

c. Frequency Multipliers. As can be seen from Table I, the current content of the bunched beam is high, even at very high order harmonics. With the high Q resonant cavities, there is no difficulty in separating the harmonics. Frequency multiplication of 10:1 is found in commercially available tubes. Experimental tubes with multiplication of 20:1 and higher have been built.

Because of paramount importance of the various effects which are of only second order in the ordinary klystron, it is found that the operation of frequency multipliers is not nearly as efficient as Table I predicts. Practical tubes with 10:1 multiplication do not have efficiencies above 1%. Very little effort has been devoted to this particular type of klystron in the past, and it is not known what degree of improvement can be expected in efficiency in the future.

On the other hand, frequency multipliers (at low efficiency) operate satisfactorily and reliably. They can be used as crystal-controlled local oscillators and when followed by a cascade amplifier permit construction of simple crystal-controlled transmitters.

d. Cascade Frequency Multipliers. Shown in Fig. 2*d* is a three-cavity klystron, the first two of which form an amplifier and the third, a harmonic output cavity for frequency multiplier use. Various combinations of cavities can be used for a number of practical purposes. It is found that the use of an intermediate cavity can improve the conversion efficiency by a large factor.

e. Two-Cavity Oscillator. By coupling some of the output from the second cavity to the first, a two-cavity amplifier can be converted into an oscillator. This was the original klystron oscillator and was used for various purposes until recently.

The two-cavity klystron oscillator can be frequency modulated by changing the accelerator voltage. For small changes in voltage, the modulation is linear, making the klystron a convenient device for such service as f-m transmitter, f-m radar, etc. The modulation power required is not high, but is much higher than that required by the reflex klystron.

A typical efficiency of a modern oscillator is the same as of the amplifier, i.e., about 30%.

f. Electron-Coupled Oscillator. This klystron, sometimes known as the "buffer type," is shown in Fig. 2*f*. It consists of a two-cavity klystron oscillator, and a third cavity immediately adjacent to the second. Since there is no drift space between the second and third cavities, the third cavity extracts the power from the beam as if it were the second. The useful load is connected to the third cavity.

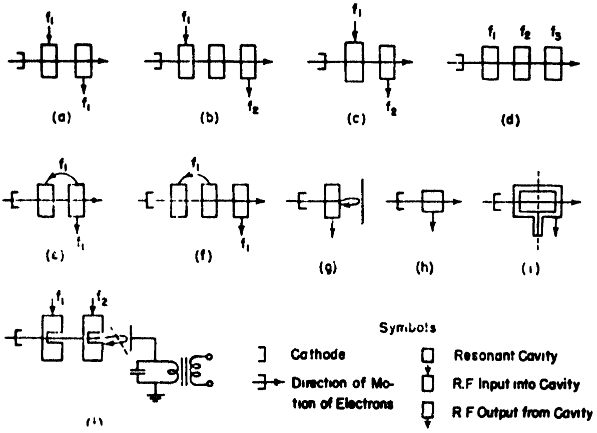


FIG. 2 Schematic representation of the various types of klystrons.

This arrangement, in principle, makes the frequency of oscillation entirely independent of load variations. In practice, it is found that the isolation is not perfect, owing to electrostatic coupling between the two resonators, and because high-speed secondary electrons from the collector traverse the tube in the opposite direction.

However, with proper precautions, a large degree of isolation is possible.

g. Reflex Klystron. This klystron, shown in Fig. 2*g*, combines the input and output interaction gaps in a single cavity. The electron beam, in passing through the cavity, undergoes velocity variations as in a two-cavity klystron. The beam then passes into a retarding region, is brought to rest in front of a reflector, and returns to the cavity. In the process of traversing the retarding region twice, velocity grouping takes place, much like that in an ordinary klystron. If the phase of the returning electrons is correct, oscillation can exist.

The reflex oscillator cannot be as efficient as a two-cavity klystron because the same cavity acts as a buncher and catcher, and the same voltage across the gap cannot be simultaneously best for the dual function.

The relatively low efficiency of the reflex klystron is not important for many applications, as for example, local oscillator service in receivers, and for low-power transmitters. The tube, having but a single cavity, can be tuned easily; the reflector electrode is highly negative, draws no current, and can easily vary the frequency oscillation over a range which is wide enough for frequency modulation applications.

The reflex klystron is the only type of klystron that has been manufactured in large quantities up to now. It has been made in a large variety of styles, methods of tuning, etc., and models are available which will cover the entire microwave frequency spectrum.

In recent years, a number of models have been developed specifically for transmitting applications. These are usually narrow-tuning tubes, with high-power input and high output.

h. The Monotron. A single cavity, with large transit time, such as $1\frac{1}{4}$ cycles, can be made to oscillate. It has been shown by Hansen[1] that such an oscillator can be understood in terms of conventional klystron theory. The electrons upon entering the cavity are subjected to time-varying electrostatic fields which cause velocity modulation. In passing through the cavity, velocity grouping takes place, and upon exit, final interaction with fields results in a net transfer of energy to the cavity.

The monotron, in essence, is a two-cavity klystron of the Heil type (see *i* below) in which the drift tube has been left out. The time-varying forces which act upon the beam during the bunching time merely modify the detailed process.

Since the cavity is not reentrant, it is larger than that of a conventional klystron and is suitable for high-power operation. Although the elementary theory of Hansen has been verified by Alpert,[13] no successful practical tubes have been reported.

i. The Floating Drift-Tube Klystron. Known in Europe as the Heil tube, after Heil's description,[14,15] the klystron shown in Fig. 2*h* can be understood in terms of an ordinary klystron. If the fields in the buncher and catcher are equal and of opposite phase, then the partition between them can be left out without disturbing the fields in the cavity. If the drift tube is supported by $\lambda/4$ support, the action of the klystron is not disturbed.

Tubes of this type have been studied in detail and have been built in many different geometrical forms. Some of these can be tuned by changing the length of the drift-tube support, and result in wide-range, single-control oscillators.

The drift tube of such a klystron can be insulated for direct current, and the frequency of the klystron can be varied for frequency-modulation purposes by varying this potential. The behavior of this type of klystron is analogous to the reflex tube, except that the power-handling capacity is much greater.

j. Klystron Converter. A klystron can be used as a frequency converter in several different ways. None of these proved to be sufficiently low noise to be practical in the past, and none has been made commercially. If, however, successful cascade amplifiers are developed for receiver use, the high noise level of the converter can be overcome by sufficient preamplification. In this event, the simplicity and reliability of a klystron type tube may prove to be of interest to warrant further development.

An example of klystron type frequency converter is shown in Fig. 2*j*. The usual arrangement of cavities is followed by a grid and a collector plate, the ensemble forming an electrostatic velocity sorter. The grid is connected to the cathode and the plate to the input terminals of an intermediate frequency amplifier, and thence to ground. The signal may be introduced into the first cavity, and the local oscillator into the second. The nonlinear action of the velocity sorter elements produces frequency conversion.

III. THEORY OF THE KLYSTRON

1. *Introduction*

Simultaneously with the invention of the klystron it was possible to give a mathematical theory of its behavior which explained most of the gross features of its operation.[9,10] Indeed, one was able to get even semiquantitative results out of this initial theory which can be described very simply.

An electron beam corresponding to a direct current I_0 and accelerated by a voltage V_0 passes through an infinitely narrow gap across which there is a voltage $V_1 \sin \omega t$. After passing through the gap the electron velocity is given by

$$u = \sqrt{\frac{2eV_0}{m}}\left(1 + \frac{V_1}{V_0}\sin \omega t\right)^{1/2} \tag{1}$$

where e and m are charge and mass of an electron, respectively.

For $V_1/V_0 < 1$, this is approximately equal to

$$u = u_0\left(1 + \frac{\alpha_1}{2}\sin \omega t\right) \tag{2}$$

where

$$u_0 = \left(\frac{2eV_0}{m}\right)^{1/2} \quad \text{and} \quad \alpha_1 = \frac{V_1}{V_0}$$

As the beam moves along the drift tube the velocity variation will be converted into density variation, and the current at any point can be calculated from the kinematics of the electrons. The amplitude of fundamental component of current at a distance z along the drift tube is given by $2I_0J_1(x)$, where $x = (\omega z/u_0)(\alpha_1/2)$ and is known as the bunching parameter. If this current passes through a second resonator with shunt impedance R, a voltage $V_2 = 2I_0J_1(x)R$ will be produced and the power delivered by the bunched beam to the second resonator will be $V_2^2/2R$. Although this theory does agree fairly well with the experimental results, it is oversimplified and omits many important factors. During the past few years a much more nearly complete and realistic theory has been developed with results which are in much closer agreement with measured properties of klystrons. This extensive theoretical development has been possible because of the favorable geometry of the klystron which permits separation of the various processes which occur in the tube in a much simpler fashion than in most microwave tubes.

Briefly, what is required for an adequate understanding of the klystron is:

(1) Theoretical analysis of beam control, i.e., methods of producing a high-density electron beam and, by either electrostatic or magnetic means, projecting this dense beam down a long, narrow drift tube.

(2) A theory of the interaction of the electrons with a radio-frequency gap which may be finite in extent and may have one of several different configurations commonly used in klystrons. What is desired from such a theory is the velocity modulation of the electron after leaving the gap, the exchange of energy between electrons and resonator (so-called beam loading), as well as such miscellaneous effects as resonator detuning and radio-frequency focusing of the beam.

(3) For all multi-resonator klystrons, that is multiplier, amplifiers, two-resonator oscillators, etc., it is necessary to know the effect of the radio-frequency fields due to space charge on the bunching process in the drift tube. There are effects of a different sort due to space charge in the reflector region of reflex tubes which are also important.

(4) A large signal theory which goes beyond the essentially small signal theory listed under (2). This is particularly necessary to predict the efficiency of the klystron under the conditions of large signals in the input gap and very large signals in the output gap. By very large, is meant radio-frequency voltages large enough to bring the electrons practically to rest.

(5) A theory of complex bunching systems. By complex bunching systems is meant any configuration of cavity or cavities such that the resultant modulation and radio-frequency current at the output gap is different from that produced by a single, narrow buncher gap. This includes the single buncher gap with transit time larger than one cycle. As will be shown, the complex bunching systems exhibit characteristics which indicate larger efficiencies.

All the problems listed here have been investigated extensively by analytic or graphical methods, and the results obtained are such that it is now possible to predict the properties of a given klystron design with good accuracy. In particular, one can predict efficiencies of a power amplifier klystron within a few per cent. It is not possible in the space of such a brief article to give all the details of these theories, but an attempt will be made to indicate the general procedures and summarize the results obtained.

2. *Focusing and Beam Formation*

Electron beams used in klystrons are of relatively high current density as compared to other commonly used beam devices such as the cathode ray tubes and electron microscopes. Consequently, the principal problems involved in the design of electron guns, focusing structures, etc., are concerned not so much with electron optics in the conventional sense (i.e., lenses and aberrations) but rather with methods of projecting and focusing a beam so as to compensate for its own space charge forces.

The principal contribution to electron gun design has been made by Pierce.[16] He showed that it was possible to design a focusing structure in an electrolytic tank so that electrons emitted from a spherical cap, a cylindrical segment, a plane disk, or a strip would travel in a straight line toward the accelerating anode. Essentially, his method consists of choosing electrode geometry in such a way that the potential conditions at the edge of the electron beam are the same as if the beam were a part of the space charge limited electron flow between infinite planes, concentric cylinders, or spheres. With this matching of boundary conditions, the electrons in a limited beam move rectilinearly. The use of Pierce type of cathode structure enables one to produce a convergent beam and project it through a small aperture in the accelerating anode. In the drift tube such a beam, initially convergent, will reach some minimum diameter and then start to spread because of space charge. If the drift tube is long enough some of the current will be lost on the walls. For some types of klystrons and for limited ranges of voltage and current, it is possible to use drift tube dimensions so that the beam will get through the tube using only electrostatic focusing as described. Also, for many practical geometries positive ions will be formed by electron collisions with the residual gas in the tube. These positive ions will form a core which neutralizes the electronic space charge and prevents spreading of the electron beam. Under some circumstances,[17] however, either too few of the ions are formed or they are drained out of the drift tube by the cathode fields at the entrance aperture, and little positive ion neutralization occurs. If neither positive ions nor the original electrostatic focusing by the gun is sufficient to get the beam through the drift tube, then magnetic focusing is required Also, it may happen that even though an unmodulated beam can be focused without loss through a drift tube, modulation will produce radio-frequency space charge effects (sometimes known as transverse debunching) which will require a restraining magnetic field to prevent current loss to the walls. An axial magnetic field can be so chosen as to provide radial forces which counteract the space charge forces and prevent spreading of the beam. The theory of such focusing has been treated in detail by various authors,[18,19] and the design of such a focusing field is quite straightforward.

3. *Theory of Gap Interaction*

The simple and highly idealized theory assumes that radio-frequency modulation gaps are of zero width corresponding to zero transit time of the electron. It also assumes plain parallel geometry with no variation of the field across the gap. Obviously this is not true of real gaps, which may consist of two plane parallel grids a finite distance apart, or of two cylindrical apertures facing each other, or a similar geometrical figuration. Typical geometries are shown in Fig. 3. For such gaps one wishes to find: (*a*) the velocity modulation produced by a given radio-frequency voltage; (*b*) the exchange of energy between a direct current beam and the gap - this is known as the beam loading and may have both a real and a reactive component; (*c*) if the current passing through the gap has a radio-frequency component at injection the delivery of energy from this beam to the circuit connected across the gap. For small voltages, it is possible to calculate all these approximately by analytical means. In practice it turns out that the approximation is quite good for most cases except that of a high-efficiency output gap in which the electrons are brought almost to rest by the radio-frequency voltage. In such a case the approximation of a small voltage is obviously not correct.

For a general gap the equation of motion of the electron in the direction of its original velocity is

$$m\ddot{z} = eE(z,\phi,r)\sin\omega t \qquad (3)$$

where z is the coordinate in the direction of motion, ϕ and r are conventional polar coordinates in the plane transverse to z, and ω and t are the modulating frequency and the time.

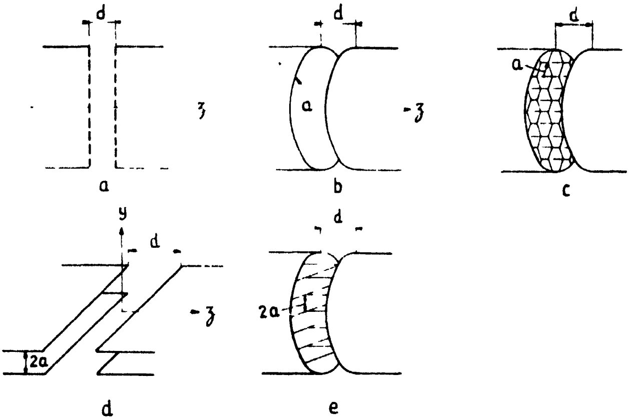


FIG. 3. Typical gap configurations. (*a*) Schematic representation of a gap; (*b*) two cylinders, or "gridless" gap; (*c*) gap with grids; (*d*) gap formed by two linear slots facing each other; (*e*) cylindrical gap with slot grids.

This problem was first solved by Petrie, Strachey, and Wallis and extended by Guenard[21,22] and Feenberg.[23] In the notation used by the first-named authors, the general equation of motion can be written as

$$\frac{d^2Z}{d\theta^2} = \frac{\alpha}{2} G(Z,R) \sin \theta \tag{4}$$

where Z, R, θ are normalized distances and times, i.e., $\frac{\omega z}{u_0}$, $\frac{\omega r}{u_0}$, ωt, respectively, where u_0 is the velocity of the unmodulated beam, and G is a normalized field given by

$$G = \frac{E(Z,R)}{\int_{-\infty}^{\infty} E(Z,R)dZ} \tag{5}$$

and

$$\alpha = \frac{1}{V_0} \int_{-\infty}^{\infty} Edz \tag{6}$$

Let θ_0 be the time at which an unmodulated electron would cross the center of the gap ($Z = 0$). Then for small voltages the approximate relation between position and time is $\theta = \theta_0 + Z$. This is the exact relation in the absence of radio-frequency fields. Obviously, this introduces a small error in the position of the electron as a function of time and, therefore, in the acceleration. However, this does give the acceleration and velocity correctly to linear terms in α, the radio-frequency voltage.

It can be seen from eq. 4 that a first order correction to the position will contribute only a quadratic correction in the acceleration. By integrating the simpler equation, one can obtain the first order correction to the position of the electron due to the radio-frequency voltage. Using this first order correction in the position, one can get the velocity of the electron at any point in its path correctly to quadratic terms in the radio-frequency voltage. These quadratic terms are used to evaluate the energy exchange between beam and the fields in the gap. As a result of this analysis, the velocity of the electron upon emergence from the gap is obtained correctly to linear terms in the radio-frequency voltage, and the energy of the electron on emergence, correctly to quadratic terms. The difference of the average energy between entrance and emergence determines the transfer of energy between the beam and the gap, i.e., the "beam loading."*

These results for the principal geometries of interest can be summarized by two parameters. These are:

(1) A beam coupling coefficient β which is a multiplicative constant representing the reduction in the velocity modulation produced by a given gap geometry as compared to an infinitely narrow gap. Equation 2 becomes

$$u = u_0\left(1 + \beta\frac{\alpha_1}{2}\sin\omega t\right) \tag{2'}$$

with a corresponding change in the definition of the bunching parameter

$$x = \frac{\omega z}{u_0}\frac{\beta\alpha}{2}$$

(2) A beam loading conductance G_b, representing the gap loading produced by the electrons. This conductance multiplied by the radio-frequency voltage, squared, equals the power lost (or gained) from the beam.

For a plane parallel gap of width d (shown in Fig. 3a) neither of these parameters will depend on the radius since the field is uniform everywhere across the cross section of the gap. For this geometry the beam coupling coefficient is

---

* By an extension of this analysis,[23] it has also been possible to find the detuning of the cavity produced by the presence of the electron beam. However, this is not of as great interest as the loading.

$$\beta = \frac{\sin \frac{D}{2}}{\frac{D}{2}} \tag{7}$$

and the beam loading conductance is

$$\frac{G_b}{G_0} = \frac{1}{2} \frac{\sin \frac{D}{2}}{\frac{D}{2}} \left(\frac{\sin \frac{D}{2}}{\frac{D}{2}} - \cos \frac{D}{2} \right) \tag{8}$$

where $D = \omega d/u_0$ is the transit angle across the gap and $G_0 = I_0/V_0$. Shown plotted in Fig 4 are curves of the beam coupling coefficient and beam loading conductance for this case.

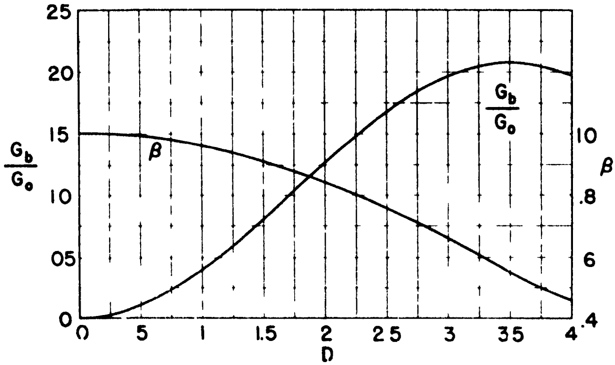


FIG. 4. Beam coupling coefficient and beam loading conductance for plane parallel gap geometry

For a gap consisting of a pair of opposed cylinders of radius a and spacing d (Fig 3b), the beam coupling coefficient depends on the radial position of the electron and is given by

$$\beta_R = \frac{I_0(R)}{I_0(A)} \cdot \beta_A \tag{9}$$

where $R = \omega r/u_0$, $A = \omega a/u_0$, and $I_0(R)$ is the modified Bessel function. β_A is the beam coupling coefficient at the edge of the gap and will depend on the nature of the field at the edge of the gap. This field, in general, is not known. One commonly assumes that the field is either (*a*) uniform, as in plane parallel geometry, or (*b*) like that between two parallel knife edges. In the first case

$$\beta_A = \frac{\sin \frac{D}{2}}{\frac{D}{2}} \tag{10}$$

in the second

$$\beta_A = J_0\left(\frac{D}{2}\right) \tag{11}$$

The true fields probably lie somewhere between these two extremes. Actually, these two alternative values of β_A do not differ greatly.

The beam loading conductance will depend on the radius of the cylinder and the radius of the beam b (assuming a uniform current density) and is given by ($B = \omega b/u_0$)

$$\frac{G_b}{G_0} = \frac{1}{2}\beta_A^2\left[\frac{AI_1(A)(I_0^2(B) - I_1^2(B))}{I_0^3(A)} - \frac{I_1^2(B)}{I_0^2(A)}\right] + \frac{1}{2}\beta_A\gamma_A\frac{I_0^2(B) - I_1^2(B)}{I_0^2(A)} \tag{12}$$

where β_A is as defined previously, and

$$\gamma_A = \frac{\sin \frac{D}{2}}{\frac{D}{2}} - \cos \frac{D}{2} \tag{13}$$

or

$$\gamma_A = \frac{D}{2} J_1\left(\frac{D}{2}\right) \tag{14}$$

depending, as previously, on the choice of the field at the edge of the gap.

For a gap consisting of two linear slots facing each other (Fig. 3*d*), of width $2a$ and spacing d, the beam coupling coefficient depends on the transverse position Y of the electron and is given by

$$\beta_Y = \beta_A \frac{\cosh Y}{\cosh A} \tag{15}$$

where

$$Y = \frac{\omega y}{u_0}, \quad A = \frac{\omega a}{u_0} \tag{16}$$

and β_A is the coupling coefficient at the edge. The beam loading conductance will depend on the width of the slots and the width of the beam b. It is given by

$$\frac{G_b}{G_0} = \frac{1}{4}\frac{\beta_A{}^2}{\cosh^2 A}\left[A \tanh A\left(1 + \frac{\sinh 2B}{2B}\right) + \frac{1}{2}\left(\frac{\sinh 2B}{2B} - \cosh 2B\right)\right] + \frac{\beta_A\gamma_A}{4\cosh^2 A}\left(1 + \frac{\sinh 2B}{2B}\right) \quad (17)$$

where β_A, γ_A have the same significance as before.

These equations are sufficient to describe completely the interaction of the electron beam with an *input gap* or a *small signal output gap*. All the relevant information is contained in these coefficients. The range of validity turns out to be quite large and can be determined approximately by considering the higher order terms which have been neglected.

It might be pointed out that all these calculations assume that the beams have no transverse components of velocity. For all except plane parallel gaps there will be transverse field components. Unless there is a restraining magnetic field, these will produce transverse velocities and, therefore, a contribution to the electronic beam loading. The latter has been calculated by Feenberg[23] for the geometries described here and is, in general, somewhat smaller than the longitudinal term.

The transverse fields also add another effect. This is an alternate focusing and defocusing of the beam due to the radio-frequency transverse velocity. An analysis of this shows that the center of the bunch is defocused, whereas the center of the anti-bunch is focused. Exact values have been calculated and can be taken into account in any klystron design.[21,23]

4. *Space Charge Effects*

The second problem to be considered is the effect of space charge in the drift tube. When a velocity modulated beam enters a drift tube and the bunching process starts, the increased density in certain regions of the beam produces electrostatic forces which tend to alter the motion of the electrons. It is quite obvious that these forces necessarily will be such as to oppose concentration of charge, and therefore inhibit the bunching process. A detailed analysis of these effects is possible by means of the method of space charge waves first introduced by Hahn[12] and Ramo[24] and developed further by Feenberg[23,25] and by Warnecke, Guénard, and their collaborators.[26,27,28] In essence, one tries to find functions which describe the charge density, voltage distribution, and velocity distribution throughout the beam which simultaneously satisfy Maxwell's equations and the equation of motion. In solving these equations, it is necessary to make some approximations. In particular, one obtains linear equations by omitting certain terms from the exact equation. This limits the range of validity of the results.

In the exact equations the nonlinear term in the alternating current comes from the product of the radio-frequency charge density and the radio-frequency velocity. If one assumes that the cross product term is small compared with the other two terms in the current, namely the product of radio-frequency charge density and direct current velocity and of radio-frequency velocity and the direct current charge density, then the resulting equations are linear in the amplitudes of field, potential, current, velocity, displacement, etc. Under these circumstances, it is possible to find solutions for these quantities for the geometries commonly used for drift tubes, i.e., circular cylinders or plain parallel boundaries, in terms of so-called space charge waves which propagate through the beam. The propagation constants will depend on the direct current beam velocity, beam density, and drift tube dimensions. For either of the geometries, there is an infinite set of such waves having different transverse dependence. For example, in the case of a cylindrical geometry, the transverse dependence is that of a Bessel function of the radial coordinate. The coefficient of the radius r in the argument of the Bessel function is determined by boundary conditions at the cylindrical boundary. These waves and their corresponding frequencies are closely related to the plasma oscillation which is characteristic of an infinite plasma and whose frequency depends only on the charge density. By the introduction of metal boundaries, however, the simple plasma calculation is modified and one finds a whole set of modes.

Each of these space charge waves will satisfy the boundary conditions at the walls. It is then necessary to take suitable combinations of individual waves to satisfy the boundary conditions at the two ends of the drift tube. These are usually conditions imposed on fields, initial velocity, and on current density. Using this basic approach of propagating space charge waves in the beam, it has been possible to calculate all the space charge effects for a number of geometries and conditions. These include cylindrical and plane parallel drift tubes, infinite and zero magnetic fields, grid coupling, and gap coupling at the input to the drift tube.

The calculations are such that they are valid only for a certain range of parameters. First it is assumed that the plasma frequency is much less than the modulating frequency. In practice, this is usually satisfactory. Second, it is necessary that none of the electron trajectories cross, i.e., that no electrons pass each other along the drift tube. In terms of the purely kinematic theory, this corresponds to values of the bunching parameter $x = \pi N\alpha$ less than 1. This is a more serious limitation than the previous one. Strictly speaking, this limits the applicability of the theory to this range of operation. Unfortunately, there is no known technique available at present of solving corresponding problems if the trajectories cross. This is the well-known problem of a multivalued stream, i.e., in which the velocity at a given point does not have a unique value. There has been little, if any, work done on this problem to date. However, the theoretical calculation for noncrossing trajectories is still of considerable importance in evaluating the performance of the klystron. It enables one to determine fairly accurately the effects of the space charge for $x < 1$ and to get some qualitative ideas beyond this.

Qualitatively, the effect of the space charge is to slow up the bunching process because of the space charge forces; eventually, if the drift tube is long enough, the bunching will stop, and the relative velocity of the electrons is reduced to zero. This will correspond to the maximum density obtainable in the beam. Beyond this point in the drift tube the density will decrease. Thus there is an optimum length of drift tube.

The most important results of the theory are the determination of this optimum length and of the numerical factor by which the current as calculated by simple kinematic theory is modified by the effects of the space charge. For a beam with infinite cross section, there are only two space charge waves and the modification of the kinematically calculated current is obtained simply by replacing the distance along the drift tube z by a factor $\sin hz\ h$ where

$$h = \left(\frac{e\rho_0}{\epsilon_0 m}\right)^{1/2} \frac{1}{u_0} \text{ in rationalized MKS units.}$$

For an infinite beam, or in practice for a beam of a large cross section, there is an optimum distance L determined by

$$hL = \frac{\pi}{2} \qquad (18)$$

Beyond this distance the bunching parameter begins to decrease. In Figs. 5 and 6 are shown the corresponding functions in the case of a cylindrical drift tube of finite diameter and with a finite beam. Here a = drift tube radius, b = beam radius

$$A = \frac{\omega a}{u_0} \qquad (19)$$

In these the value of $\langle hF \rangle$ replaces the $\sin hz$ for the simple case. The corresponding value of the length for which this reaches a maximum is the optimum drift tube length. The curves shown in the figures were calculated assuming an infinite magnetic field.

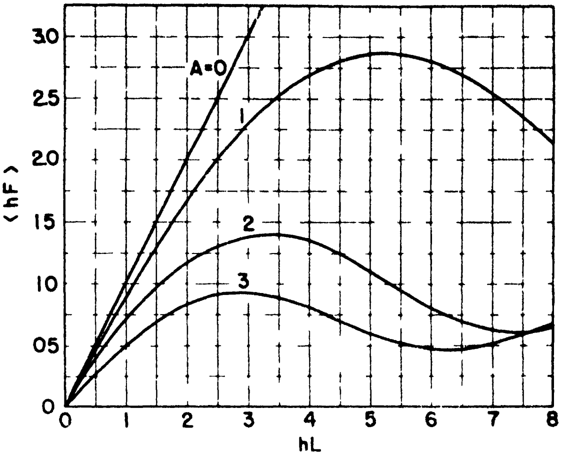


FIG. 5 Space charge reduction factor for beam filling drift tube (magnetic focusing) Curves labeled for values of $A = \omega a/u_0$

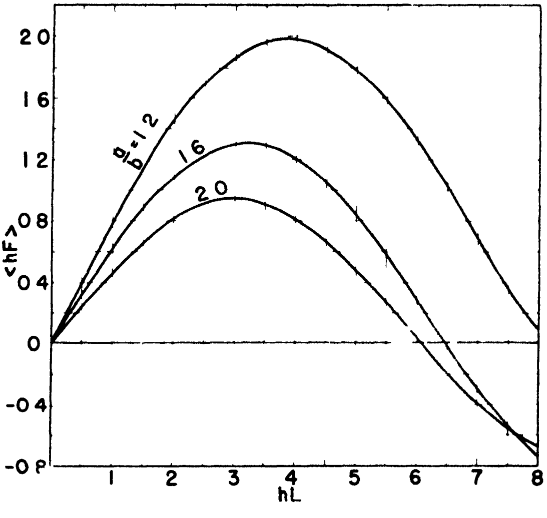


FIG. 6 Space charge reduction factor for beam partially filling drift tube (magnetic focusing) $A = \omega a/u_0 = 1$

5. *Theory of Large Signal Effects*

The calculation of large signal efficiency in klystrons is related to the general problem of the interaction of an electron stream with a radio-frequency field in a gap. However, the methods previously used to obtain analytic answers for the energy interchange between an electron beam and the gap do not apply if the radio-frequency voltage across the gap is of the same order of magnitude as the beam voltage. If the radio-frequency voltage is small, then one can obtain an effective driving current. The injected beam merely behaves like a constant-current generator driving a circuit. This method is valid only if the radio-frequency voltage is small compared to the direct-current voltage, so that it is possible to calculate the effects of the radio frequency voltage only as a small perturbation on the total motion. In the output gap of a high-efficiency klystron, this is not true. In this case, the radio-frequency voltage is almost equal to the direct-current voltage, a large number of electrons are brought almost to rest and transfer a large fraction of their total energy to the radio-frequency field. It is obvious that under these conditions, treating the radio-frequency field as a small perturbation is not valid. However, it is possible to calculate the efficiency even in such cases by graphical methods.

The methods are simple in principle, though somewhat laborious. One merely calculates electron trajectories through the buncher gap, along the drift tube to the catcher and then through the catcher gap, obtaining the final exit velocity for the given values of the gap voltages. In this calculation it is possible to use the analytic theory for the velocity modulation on leaving the first gap and also for the motion along the drift tube. The motion through the catcher gap must be computed numerically This process is repeated for a number of electrons with entrance phases at the buncher gap equally spaced over a cycle. The average value of the kinetic energy of the electrons after passing through the output gap is compared with their average energy on entrance to the gap. The difference gives the transfer of power. The number of trajectories which must be averaged depends on the accuracy desired. For nonplanar gaps, in addition to the process of averaging over a cycle, it is necessary to average across the cross section, since electrons at different transverse positions will have a different motion in the gap. This type of calculation has been carried through in some detail by Feenberg[13] and by Warnecke and collaborators[29] for plane gaps. The results are shown in Fig. 7. The normalized current, $F_c = I_c/2I_0$, driving the catcher is plotted against the relative voltage (α_2) across the catcher, for various values of the transit angle (D_2) across the catcher and for various values of the buncher voltage (α_1). The efficiency is given by $\alpha_2 F_e$. For zero width gaps ($D_2 = 0$) F_e is a constant up to the point $\alpha_2 = 1 - \alpha_1$ where electrons begin to be turned back by the catcher field. This accounts for the sudden break in the $D_2 = 0$ curves and also the drooping of the $D_2 = 1$ curves for large α_2 For small α_2 the F_e curves vary according to $F_e = 0.58\beta - \alpha_2 G_b/2G_0$ as predicted by the analytic

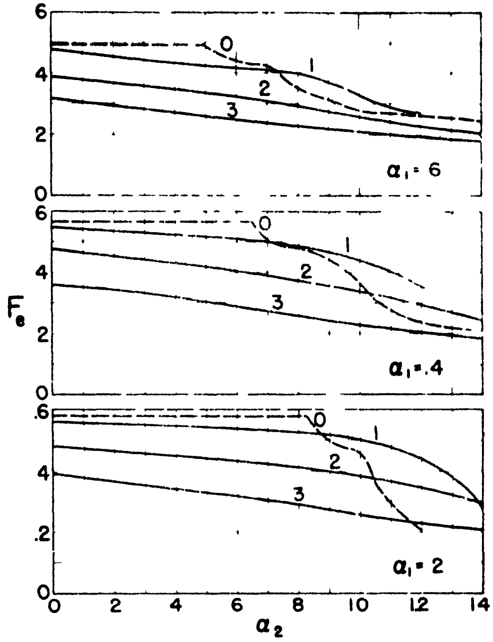


FIG. 7 Large signal conversion efficiency. α_2 is the ratio of the radio-frequency voltage at output gap to beam voltage, α_1 is this ratio for the input gap. The ordinate I_e, is the effective driving current and the conversion efficiency is $\alpha_2 I_e$. The curves are labeled with the values of the direct-current transit angle across the output gap, D_2. α_1 is the *effective* voltage at the input gap and contains implicitly the input transit angle.

small signal theory for finite gaps It should be noted that for finite catcher gaps and large catcher voltage the efficiency is reduced below the theoretical 58% value calculated in the very simplest theory.

The results shown in Fig. 7 represent total conversion efficiency, i.e., the power delivered from the beam to the cavity and includes power delivered to the load as well as cavity losses. By changing the loading of the output cavity, it is possible to change the resonator voltage and it is also possible to measure this relative voltage. From such a combination of measurements it is possible to determine the total conversion efficiency and it has been found to be in fair agreement with the theory indicated here.

6. Theory of Complex Bunching Systems

The last major field of theoretical development involves the study of structures which attain larger efficiencies than the conventional klystron, and/or larger gains (for amplifiers). These can be grouped under the title of complex bunching systems. One of these structures will be discussed here briefly, namely, a klystron amplifier, with three resonators instead of two.

The first resonator driven with a low power produces a small current at the second resonator. But because of the high impedance of the second cavity, it can produce a relatively larger voltage across this second buncher. If the parameters involved are correctly chosen, the voltage across the second cavity will produce a much larger amount of modulation than that due to the first cavity. The effect of the first cavity at the output gap can almost be neglected. In this method of operation, the tube is merely a high-gain amplifier and its efficiency and all other properties are similar to those of a two-resonator amplifier.

However, if there is a large radio-frequency voltage across the first cavity, the current at the second cavity can be appreciable and under normal circumstances would produce a very large radio-frequency voltage perhaps of the order of the beam voltage. However, by suitable detuning or loading of the second cavity this radio-frequency voltage can be reduced to about the same value as that across the first cavity. Two things must now be pointed out. First, when the beam arrives at the second cavity, it is already partially bunched. Second, modulation produced by the voltage of the second cavity is not in phase with that produced by the first. This relative phase can be controlled by the amount and direction of detuning of the second cavity. The net result of the two modulations can be such as to produce more effective bunching with a larger fundamental component of current at the catcher. Figure 8 shows the arrival times of the electrons at the catcher if velocity modulation is produced by only one gap and also when produced by two modulating gaps suitably spaced and tuned. In the figure, $\omega t_2 - \theta_0$ is plotted against ωt_1. ωt_1 is the departure time from the first gap, ωt_2 is the arrival time at the catcher, and θ_0 is the transit angle between initial and final gaps, for an unmodulated electron. It can be seen that most of the electrons arrive during a much shorter interval of time in the second case and, therefore, produce a larger fundamental component of current.

By a similar simplified theory, which predicts 58% efficiency for a two-cavity klystron, one finds that for a suitably adjusted three-cavity tube the efficiency is about 80%. Of course, the effects of space charge, large signals, etc., will reduce both these numbers, but the ratio is significant. A close examination of the theory indicates why this improved current is obtained from two gap modulation and also suggests alternative ways of getting this improvement.

From the theory,[23,30] it can be shown that the effect of this double gap bunching, with suitable phase adjustment of the second cavity, is equivalent to modulation by a combination of two sinusoidal voltages, the second of which is at twice the frequency of the fundamental and with the amplitudes and phases of the two components related in the same way as in the Fourier expansion of a sawtooth voltage. Thus, the two-gap modulation is a method of approximating the result obtained by sawtooth modulation. It can easily be shown that a sawtooth modulation would produce maximum possible radio-frequency current amplitude (and indeed at all harmonics).

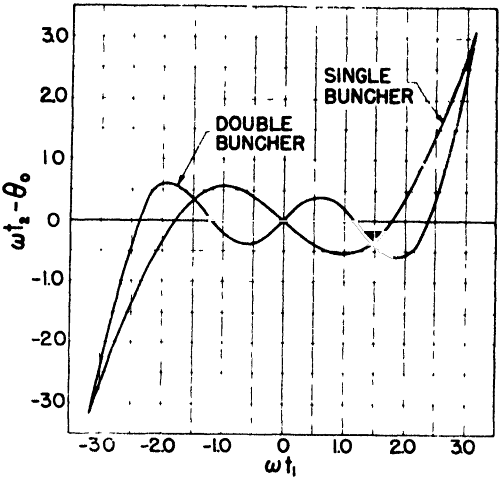


Fig. 8. Electron arrival time at catcher for optimum bunching.

An alternative way of getting the beneficial effects of a second harmonic term in the modulation has been suggested.[31] It involves using a very wide input gap and a very large radio-frequency voltage. By wide gap is meant a gap with a transit angle equal to or greater than one cycle. An analysis of the motion of the electrons through such a gap indicates they emerge with velocities like those produced by modulation by the first and second harmonic terms of a sawtooth voltage. Therefore, the resulting bunching also has the same characteristics as those previously described and can lead to higher efficiency. An obvious third scheme which has been suggested would be to use two bunchers back to back, one at twice the frequency of the other.

IV. KLYSTRON AMPLIFIERS

As pointed out previously, klystrons are useful for several classes of amplifier service. These will be discussed separately below; however, some remarks can be made which apply generally.

Practically all contemporary klystron amplifiers use three cavities. In the case of high-power tubes, the third cavity makes it possible to increase the efficiency or to increase power gain, whichever happens to be more important. In low-power amplifiers, the addition of an extra cavity makes it possible to build extremely high-gain tubes. In low-noise amplifiers, the addition of an extra cavity improves the noise level.

It is natural to ask, then, why not use more resonators than three? In the first place, an extra cavity adds complexity to the tube. Secondly, because of the extra length of drift space the loss of beam current to the walls of the tube becomes a serious factor. This is especially true in electrostatically focused beams. In the third place, the addition of extra cavity or cavities is not always helpful. For example, addition of a fourth cavity does not materially change the efficiency of a high-power amplifier. For such applications as voltage amplifiers, extra cavities would be of value in increasing the gain. The limit to the number of cavities would be determined by the regenerative effects due to fast secondary electrons returning to the cathode from the collector. In practice, it has been found that a six-cavity tube with 60 db of gain could not be operated without oscillation.

Another important distinction between various klystron amplifiers is in the construction of the cavity interaction gaps. There are two ways in which cavity gaps can be made, as shown in Fig. 3. Figure 3*b* indicates a gap formed by two cylinders facing each other. Figure 3*c* has an electron permeable membrane, called the "grid." A grid must not intercept a large fraction of the beam. If it does, the loss in current reduces the efficiency. In high-power tubes this causes overheating of the grids.

The gap with grids prevents fringing of the gap field. In this case the gap spacing is the only parameter that affects the beam coupling coefficient. Then the grid diameter can be chosen as an independent quantity. In particular, in beams with large diameter, low-current density can be used; and the beam spreading problem is relatively unimportant.

In the case of the gridless gap, the coupling coefficient depends upon both the gap spacing and the diameter of the gap. Because of this, the diameter of the gap depends upon the beam voltage and, for the usual voltages used in klystrons, the gap diameter is considerably smaller than it can be with grids. Hence, the beam focusing problem is more severe, especially in low-voltage tubes.

As a consequence of these considerations, low-voltage tubes are always built with grids. But when the power input exceeds, for example, 1000 watts in a 10-cm tube, grids must be abandoned. In high-power, high-efficiency tubes, with carefully designed focusing systems, the grids are both impractical and unnecessary.

1. Low-Noise Amplifiers

Very little work has been done to develop low-noise klystron amplifiers for receiver applications. Klystrons, designed for power service, exhibit noise figures of 25 40 db.* The mechanism of noise generation, the effect of space charge smoothing and the variation of noise along the beam due to noise velocity modulation at the cathode are factors which are still insufficiently understood to allow one to design tubes intelligently. However, during the last three years, experiments at Stanford University and elsewhere have indicated that space charge smoothing is almost as effective at microwave frequencies as it is at the lower ones. Much of the noise generated in present-day klystrons is due to defects that can be corrected: an example is the ordinary interception noise caused by grids.

It is believed that klystron amplifiers (with several cavities) can be built to have a noise figure of about 10 db at 10 cm by careful design of the cathode gun and an optimum choice of other parameters. The gain of such a klystron can be moderately high, such as 20 db, but its bandwidth will be in the order of a few megacycles. In applications where simple mechanical structures are desirable and where the performance such as given above is of interest, the klystron may find eventual use. At this date, however, there does not seem to be any special advantage of the klystron over the wide-band traveling wave tube, where noise figures of approximately 10 db have already been demonstrated.

2. Low-Power Amplifiers

There are many applications where high amplification is desirable and the noise problems are of secondary importance. An example is an intermediate stage in a crystal-controlled transmitter. For such use, cascade klystron amplifiers are ideally suited.

---

* Noise figure expresses the ratio of receiver noise, referred to the input terminals, to $kT\Delta f$, where k is Boltzmann's constant, T is room temperature in degrees centigrade, and Δf is the bandwidth of the amplifying system. Superheterodyne receivers employing crystal converters can be made to have a noise figure of about 6–8 db.

Consider first the two-cavity klystron amplifier. The radio-frequency current I_2 at the second cavity is given by the expression (Webster)

$$I_2 = 2I_0J_1(x) \tag{20}$$

where I_0 = direct current beam current

$x = \pi N\left(\frac{V_1}{V_0}\right)$

N = transit time from buncher to catcher, in cycles

V_1 = peak radio-frequency voltage at buncher

V_0 = direct current beam voltage

If the bunching voltage is small compared to V_0, then $J_1(x)$ can be replaced by $x/2$, and the radio-frequency current becomes

$$I_2 = \pi N\left(\frac{I_0}{V_0}\right)V_1 \tag{21}$$

Defining transconductance g_m in the usual way,

$$\begin{aligned} g_m &= \left(\frac{I_2}{V_1}\right) \\ &= \pi N\left(\frac{I_0}{V_0}\right) \end{aligned} \tag{22}$$

For a set of typical values, such as $N = 5$, $I_0 = 10$ ma, $V_0 = 500$ volts, g_m has a value of 314 micromhos. This is not a particularly high value of transconductance. Fortunately, the shunt impedance values of cavities are high, and a value of 1×10^5 ohms at 10 cm is a typical value. Hence, voltage amplification per stage of the order of 30 is possible.

If higher amplification is desirable, two klystrons could be connected in cascade. However, it is more efficient to construct a single tube with three cavities along the same beam. In this arrangement, the density modulation at the second gap produces a voltage across the second gap. This in turn produces further velocity variation in the beam. Because of the high-voltage gain of the first stage, the resultant velocity variation at the second gap is practically entirely due to the second gap. It is possible to consider the action of the second-third gap region as if it were practically independent of the first-second gap space.

The second gap is not connected to any external load. Hence the voltage gain of the first-second gap region is higher by a factor of 2 than it would be if the second cavity were the output cavity and were matched to an external load. Thus, the voltage gain of a three-cavity cascade amplifier is twice the gain of two identical two-cavity klystrons. In addition, there is one beam instead of two, and three cavities to tune instead of four.

The high gain of a cascade amplifier is obtained by operating the second cavity unloaded, and therefore with a high Q. Thus, the price paid for the high gain is the decrease in bandwidth. This may or may not be important, depending upon specific applications.

Very few klystrons have been specifically designed for high amplification service. Sperry 2K35 tube was designed for combination power amplifier-voltage amplification use. While not at all ideally suited as a voltage amplifier, it has the following approximate characteristics: at a beam voltage of 2000 volts, 80 mA. beam current, at 10 cm, the power amplification is about 2000; and the bandwidth of about 1.5 mc/sec. Much better amplifiers could be built if needed.

Six-cavity, five-stage amplifiers have been built at Stanford. Operating at about 1000 volts and 30 mA., such tubes have had power amplification of about 10^5 at 10 cm. The number of stages in such an amplifier is usually limited by the presence of returning secondary electrons from the collector.

3. Power Amplifiers

Power amplifier klystrons do not differ materially from voltage amplifiers in appearance. Noise, voltage gain, and linearity were of paramount importance in voltage amplifiers. In the tubes designed for power amplification, the main factors of interest now are: power handling capacity, efficiency, bandwidth, ease of modulation, power amplification. The relative importance of these factors obviously depends upon specific use.

The power handling capacity of the klystron is very large. This is due to the fact that the regions of electron emission, radio-frequency generation, and heat dissipation can be conveniently separated. By proper choice of geometrical configurations, the cathode area and heat dissipators can be made very large, even for a simple klystron such as shown in Fig. 1. In such a structure, large spherical cathodes can be used to focus the beam into the drift space. With the aid of suitable magnetic fields, almost the entire beam can be delivered to the collector. This can be as large as one wishes. Thus, there are no problems connected with available current nor with dissipation. The power handling capacity of a klystron at a given voltage depends almost entirely upon space charge phenomenon in the drift space. When the space charge effects become serious, or if it is desired to increase the cathode or dissipation areas, the various configurations discussed in Section II can be used.

To give some idea of possible numbers, klystrons have been built at 40 cm with a total input of 25 kw. It is believed that 250 kw would be just as practical.

The efficiency of a power amplifier klystron depends upon a large number of interrelated factors. The more important of these have been discussed in Section III, 3, 4, 5. The following expression contains the various factors which affect efficiency. For simplicity of discussion it is intended to describe the behavior of a two-cavity klystron. Similar expressions can be written for cascade amplifiers. Let η be the efficiency of a two-cavity amplifier under typical conditions of operation. Then,

$$\eta = 0.58\mathrm{A}\cdot\mathrm{B}\cdot\mathrm{C}\cdot\mathrm{D}\cdot\mathrm{E} \tag{23}$$

where the factors A,B,C,D, and E express the reduction in efficiency due to the various effects and are equal to, or less than unity. These are:

A. Large signal (kinematic) correction. Owing to the space charge debunching effects, it is necessary to use a short drift space and large bunching voltage α_1. Also in the catcher, the voltage α_2 is large and is ideally equal to V_0. In practice it is approximately equal to $V_0(1 - \alpha_2)$. These effects, together with finite transit time through the catcher have been analyzed and the resultant efficiency is shown in Fig. 7. Correction factor A is the ordinate of Fig. 7 divided by $0.58/\alpha_2$.

B. Direct current loss. In practical tubes, some of the beam is lost because of imperfect focusing, presence of grids, etc. Fractional current through the tube is the factor B.

C. Transverse debunching. Due to transverse debunching, the effective radio-frequency current may be reduced. In a good design, the drift space and beam geometry are so chosen that this effect is not important. However, if the beam fills the drift tube, the loss in current due to debunching is given by

$$C = 1 - \frac{(hl)^2}{6} \tag{24}$$

where l is the length of drift space and $h/2\pi$ is the debunching wave number. In many practical tubes, it is found that C is about 0.85.

D. Cavity losses. Not all the direct current power that becomes converted into radio-frequency power appears outside of the klystron. Some is lost in the catcher in the following ways:

(1) Ohmic losses in cavity walls.

(2) Beam loading; this effect has been included in A above.

(3) Secondary electron loading due to acceleration of secondary electrons.

(4) Multipactor loading due to synchronous oscillations of secondary electrons with the field in the cavity.

Effects (3) and (4) can be minimized by suitable choice of cavity proportions and focusing of the beam. In properly designed tubes, effect (1) is negligible. However, the combined effect of all of these can be very serious.

E. Decrease in radio-frequency current due to wall effects. The debunching forces are not uniform across the beam. At the edges of the beam, the debunching forces are at a minimum. They are greatest in the center. Thus, optimum bunching does not occur uniformly over the cross section. This effect is not important if the bunching distance is chosen to minimize normal debunching.

The factors A,B,C,D, and E are all interdependent. For example, the diameter of the drift space simultaneously affects the effective gap coupling coefficient and the current density. Consequently, this affects the debunching, the percentage of the current through the tube, and the shunt impedance of the cavities. The proper choice of the various parameters cannot be made in unique fashion, and, in fact, can be made only on the basis of practical experience.

The efficiency of practical klystron amplifiers varies with frequency. This happens because the problem of passing a beam of electrons becomes harder at shorter wavelengths and because the circuit losses increase. In continuous wave operation, a power output of 50 watts with 5% efficiency at 3 cm has been obtained. At longer wavelengths the power and efficiency increase. Twenty-five hundred watts at 20% efficiency is typical of well-designed tubes.[31] Efficiency of 40% has been reported in experimental tubes in the vicinity of 1000 mc.

V. Reflex Klystrons

A schematic drawing of a reflex klystron is shown in Fig. 2*g*. Electrons from the cathode pass through the resonator. If a radio-frequency voltage is present the beam is velocity modulated. In the region between the cavity and the resonator the electrons are brought to rest, the motion is reversed, and they return to the resonator. The transit times in the reflector region for electrons of different velocities will be different, and consequently bunching will occur. For the usual variation of potential in the reflector region, it is easy to show that the electrons which have been accelerated on the first passage through the resonator will penetrate further into the reflector region and will take longer to return. In the reflex tube, therefore, in contradistinction to the drift tube type of bunching, the slow electrons overtake the fast ones. If the average transit time is suitably adjusted by setting the reflector voltage properly, then the bunched beam current comes back through the resonator in such phase as to deliver energy to the resonator and maintain the oscillation.

The advantages of the reflex klystron and the reason for its wide application are: simplicity of tuning a single cavity; since the cavities commonly used have fairly high Q, frequency stability of these oscillators is quite good; simplicity of electrical tuning, i.e., by varying the reflector voltage. Changing the reflector voltage varies the phase of the effective driving current relative to the resonant voltage. The oscillation occurs at a frequency consonant with the circuit characteristics and this relative phase. Since the reflector draws no current and its capacity is small, it is possible to obtain frequency modulation or frequency control with no appreciable consumption of power.

The theory of the reflex klystron is quite simple,[32,33,34,35,36] and experiments have shown that it quite accurately describes all the phenomena appearing in this type of tube. Briefly, the theory can be summarized as follows. The radio-frequency voltage across the gap produces a known velocity modulation of the electron beam. From the analysis of the kinematics of the electron motion in the reflector region, it is possible to calculate the radio-frequency current returning to the gap. The simple bunching calculation for the reflector region includes only linear terms in the modulation velocity just as is commonly done for a field-free drift space. It turns out that the radio-frequency current is also represented by a Bessel function as in the drift tube case. The only difference introduced by the reflector bunching is that there is a phase shift of 180 degrees in the current as compared to field-free bunching, due to the fact that the fast electrons have a longer transit time than the slow ones.

The result of the analysis can be stated in the form of the ratio of the radio-frequency current in the returning electron beam to the gap radio-frequency voltage. This is called the electronic admittance. It is, in general, nonlinear, i.e., the ratio of the current to voltage will not be a constant. In particular, it will depend on the amplitude of the radio-frequency voltage as well as on the fixed parameters such as the transit angle in the reflector region, the direct current and the direct-current voltage. Specifically, it can be written as:

$$Y_e = \frac{-\beta i}{V} = \frac{-\beta^2\theta_0}{2} G_0 2 \frac{J_1(x)}{x} (\cos \phi - j \sin \phi) \tag{25}$$

where β = the beam coupling coefficient

$G_0 = \frac{I_0}{V_0}$

θ_0 = the transit angle in the reflector region

$x = \frac{\theta_0}{2} \frac{V}{V_0}$

$\phi = \theta_0 - 2\pi(n + \frac{3}{4})$

The radio-frequency current returning to the gap will constitute an effective current generator driving the cavity. The voltage produced across the cavity is determined in magnitude and in phase by the cavity properties and the frequency of oscillation. This ratio of current to voltage as determined by the cavity is given by

$$Y_c = \frac{\beta i}{V} = \omega C\left[\frac{1}{Q_0} + \frac{G_l}{Q_c} + j\left(2\delta + \frac{B_L}{Q_c}\right)\right] \tag{26}$$

where Q_0 = the unloaded Q of the cavity

$\frac{\omega C}{Q_0}$ = the conductance of the cavity

$\delta = \frac{\omega - \omega_0}{\omega_0}$ = the fractional departure of the operating frequency

from the resonant frequency of the unloaded cavity, and $G_l + jB_L$ is the load admittance measured at a suitable place in the transmission line to which the tube is coupled and normalized with respect to the characteristic admittance of the line.

Q_c is known as the *external Q* which measures the coupling between cavity and transmission line. Its significance may become apparent from its relation to the loaded or operating Q of the loaded tube. This is

$$\frac{1}{Q_l} = \frac{1}{Q_0} + \frac{G_l}{Q_c}$$

In the steady state these two values for the admittance, one determined by the electron bunching process and the other by the cavity, must be equal. Setting them equal determines the two equations, one from the real part and one from the imaginary part. These two equations determine completely the behavior of the reflex klystron. It is common to discuss the behavior of the reflex klystron in terms of two equations, the real equation and the ratio of the real and imaginary equations as the second. This results in

$$2\frac{J_1(x)}{x}\cos\phi = \frac{\omega C}{G_e}\frac{1}{Q_0} + \frac{G_L}{Q_c} \tag{27}$$

$$-\left(\frac{1}{Q_0} + \frac{G_l}{Q_c}\right)\tan\phi = 2\delta + \frac{B_L}{Q_c} \tag{28}$$

where $G_e = \beta^2\theta_0G_0/2$.

It has been assumed that the voltage distribution in the reflector region is linear. Owing to curvature of the reflector electrode or owing to space charge effects, the voltage distribution is often nonlinear. The analysis can be modified very simply to take this nonlinearity into account. The transit angle appears in the theory in two ways. First, in the argument of the Bessel function having to do with the bunching process. Second, in the phase angle which has to do with the phase of the radio-frequency current. The phase of the radio-frequency current actually depends on the average transit time of the electrons in the reflector region independently of the form of the reflector field. This dependence will not be altered by changing the reflector geometry or characteristics in any way. However, any reflector effects which make the potential distribution nonlinear will make the magnitude of the bunched current depend on the transit time in a different way than predicted by kinematic theory.

It can be shown quite generally[3] that the argument of the Bessel function for an arbitrary reflector field should be given by

$$x = V\left(\frac{d\theta}{dV}\right)_{V=0} = \frac{\theta_0}{2}\frac{V}{V_0}\frac{2V_0}{\theta_0}\left(\frac{d\theta}{dV}\right)_{V=0} \tag{29}$$

where $(d\theta/dV)_{V=0}$ is the change in transit time of the electrons with radio-frequency voltage. For a linear reflector field $(d\theta/dV)(2V_0/\theta_0) = 1$, and x reduces to the previously stated value. For a reflector field which is not linear, one must evaluate $d\theta/dV$. The transit time θ_0 as used in the simple theory must be replaced by the effective value $2V_0 (d\theta/dV)$ everywhere except in the phase. However, the evaluation of $d\theta/dV$ for a particular reflector geometry or space charge condition may be difficult Even if a specific determination of effective transit time is not possible, it should be recognized that the theory does contain this factor and that it must be considered in any attempted correlation of reflex tube performance

Equation 27 determines the amplitude of oscillation as a function of all the operating parameters, including the reflector voltage Varying the reflector voltage changes the transit angle in the reflector region and, therefore, the phase ϕ of the radio-frequency current. Equation 28 specifies how the frequency of oscillation varies with change of transit angle. These variations of frequency and amplitude with reflector voltage lead to the typical appearance of a mode, characteristic of reflex klystrons. This is shown in Fig. 9. As the cos ϕ is varied from its optimum value, then for oscillation to be maintained $2J_1(x)/x$ must increase. The values of reflector voltage for which $2J_1(x)/x$ has to be unity determines the edges of the mode. For values of the reflector voltage beyond this, oscillation cannot exist.

In applications of the tube the two quantities of greatest interest are the optimum power for a given transit angle and the frequency width between the half-power points on a mode. To find the optimum power delivered to the load one can write load efficiency as

$$\eta_L = \frac{x^2}{\theta_0}\frac{\omega C}{G_e}\frac{G_L}{Q_c} \tag{30}$$

To get the optimum value of η_L and the corresponding values of x and G_L it is necessary to differentiate this with respect to x using the relation between x and the load admittance given by eq. 27. After some manipulation, it can be shown that at optimum power x will satisfy the relation

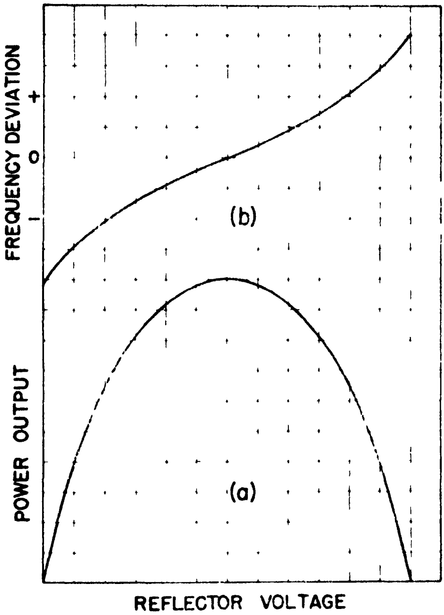


Fig. 9. Power (a) and frequency (b) variation in reflex klystron mode

$$J_0(x) = \frac{\omega C}{G_e}\frac{1}{Q_0}$$

and correspondingly

$$J_2(x) = \frac{\omega C}{G_e}\frac{G_L}{Q_c}$$

The efficiency will then be

$$\eta_L = x^2\frac{J_2(x)}{\theta_0} \tag{31}$$

Figure 10 shows a curve of relative efficiency normalized in a convenient way plotted against transit angle, also normalized in a suitable way. From this curve, one can get the efficiency for the various modes and also determine the effect of changing any of the parameters of the tube, i.e., beam voltage, beam current, and cavity properties. It can be seen from this curve that there will be a best mode, i.e., one which delivers the most power. For this case, it can be shown that power dissipated in the load will just equal the power dissipated in the cavity. There will be modes on either side of it which may not be much less efficient. A mode to the right will correspond to a shorter transit time in the reflector region and cavity losses which are larger than the load losses. On the other side of the peak, the transit angle is longer and the cavity losses are less than the load losses. For the values of load calculated by this method, it is also possible to calculate the half power $2\delta_L$ bandwidth for each such mode. This is also plotted in Fig. 10. It indicates graphically that it is possible, with some slight sacrifice in efficiency, to increase the bandwidth considerably and this is a property of considerable importance in applications where a wide electronic tuning range is important.

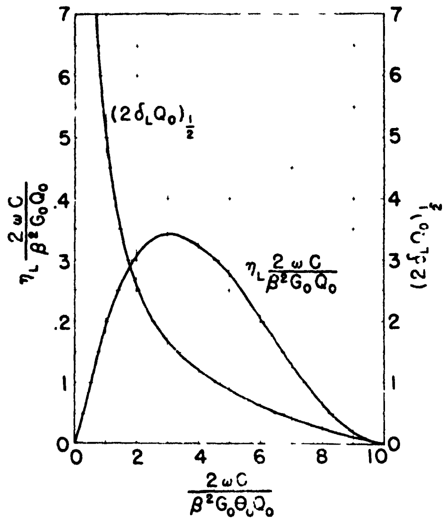


FIG. 10. Normalized efficiency and bandwidth curves.

As far as electrical tuning is concerned, another property of importance is the *modulation sensitivity* at the center of the mode, i.e., df/dV_r. This cannot be specified in terms of the cavity and beam properties alone, but will also depend on the reflector spacing. A closer reflector spacing will result in a lower reflector voltage for the same transit angle and also a higher modulation sensitivity. One can write the expression for the modulation sensitivity as

$$\frac{df}{dV_r} = \frac{1.2\,\theta_0^2}{S\sqrt{V_0}\,Q_l}\ \text{Mc volt} \tag{32}$$

where S is reflector spacing in centimeters, and Q_l is loaded Q. It is assumed here that the reflector field is linear, but this relation is probably accurate even when this condition is not exactly satisfied. This equation gives some idea of what modulation sensitivities are possible, but it must be mentioned that there is a restriction on a reflector spacing, or more strictly speaking on reflector voltage. If the voltage is too close to zero, i.e., cathode potential, then electrons which have been accelerated in the resonator gap will strike the reflector and extreme distortion of the mode will occur. Therefore the reflector spacing must be great enough to prevent this.

Hysteresis. The small signal theory discussed up to now assumes that the phase ϕ of the current depends only on the reflector voltage and, in particular, is independent of the amplitudes of oscillation. This is only true for small amplitudes and, therefore, eq. 27 determines the two values of ϕ (and V_r) at which oscillation will start, if these values of V_r are approached from the nonoscillating region outside the mode. However, for large amplitudes there are effects not included in the simple theory which may cause the phase to be a function of the amplitude. In that case it is possible that at large amplitudes the phase for some range of V_r near the edge of the mode will be shifted enough from the zero amplitude value so that oscillation is possible at this amplitude even though it cannot exist at small signals. Therefore, if the reflector voltage is varied from the oscillating region at the center of the mode through this range of V_r at the edge, oscillation will continue even though it could not start from zero amplitude if the mode were approached from the outside. The width of the mode will, therefore, depend on the direction in which V_r is swept.

Long Line Effects. Another source of hysteresis may be met with when the load admittance $G_L + jB_L$ varies rapidly with the frequency. This is particularly the case when the reflex klystron is connected to a long line not perfectly matched. The amplitude of the variation of the load admittance is determined by the reflection coefficient at the end of the line but the rate is proportional to the length of the line, giving rise to very rapid variations when the line is long. Even when the reflection coefficient at the end of the line is not large enough to give rise to hysteresis, then "long line effects" introduce a very large distortion in the frequency modulation characteristics of the tube.

Typical Characteristics. As an indication of the properties of practical reflex tubes, it might be useful to list the operating characteristics of some commercially available tubes: it has been found possible to get as high as 12 watts with a bandwidth between half-power points of about 20 megacycles at an operating frequency of 1500 megacycles; and as high as 5 watts with a bandwidth between half-power points of about 40 megacycles at an operating frequency of 7000 megacycles. The modulation sensitivity varies from about 1⁄10 megacycle per volt for the former tube to 1⁄2 to 1 megacycle per volt for the latter tube. The efficiencies obtained vary from about 7 or 8% down to 4 or 5% for the tubes mentioned. Corresponding powers, bandwidths, and efficiencies are either available or obviously possible anywhere in the frequency range between these two examples. Somewhat lower powers have been obtained at higher frequencies, and small bandwidths at lower frequencies. The tubes mentioned were designed primarily for transmitter use in microwave relays and for practical reasons did not incorporate extensive mechanical tuning ranges. For use in bench measurements, this amount of power is not required. For such purposes there are tubes available which have a tuning range of plus or minus 20%, and a power in the neighborhood of 1⁄2 to 1 watt everywhere from about 13,000 down to about 1000 megacycles. At higher frequencies the power available is somewhat less, being perhaps as little as 30 or 40 milliwatts at 30,000 megacycles. This does not necessarily represent the ultimate limit possible at these frequencies.

VI. SUMMARY

By necessity, the foregoing description of klystron theory and performance has been brief. In summarizing, it is probably most illuminating to list the most important functions which can be performed by klystrons; describe how much improvement can plausibly be expected in the light of present knowledge and performance; and finally, to make some comparisons with other types of microwave tubes which perform these same functions.

Local Oscillators and Laboratory Oscillators. Reflex tubes fall into three categories:

(*a*) The narrow tuning range, low-power output tubes used in superheterodyne receivers as local oscillators. For this service, the reflex tube is ideal and will probably not be displaced appreciably by any other device now existing. As presently constructed, these tubes are simple and lend themselves to quantity production.

(*b*) Wide-range tuning, medium-power tubes are used in laboratory oscillators, "search" receivers, and in other special applications. Klystrons of this type require complicated tuning mechanisms and may eventually find serious competition from tubes of the traveling wave type. Both the reflex klystron and the traveling wave tube have about the same low efficiency; this is seldom an important consideration.

(*c*) High-power reflex tubes. These will be discussed in a separate section below.

It is useful in many applications to have a simple method of automatic frequency control. In a reflex klystron, this can be provided by controlling the reflector voltage. This can also probably be accomplished by beam voltage control in a traveling wave oscillator. However, the latter is not as convenient as using the high impedance reflector electrode.

Laboratory oscillators have to perform various functions. Reflex tubes are very convenient because of the ease of tuning, low power supply requirements, and fairly large output

Relay Transmitters. For low power (1–10 watts) point-to-point microwave relay links, the reflex tube is useful because of its simplicity and its modulation characteristics. A traveling wave oscillator is also a possibility. However, little is known about its modulation characteristics, and for this application, its wide mechanical tuning range would be of little importance. A variation for such links might be a combination of reflex tube and traveling wave amplifier. In such a case the reflex tube could be designed for optimum modulation characteristics.

Power Amplifiers and Multi-Cavity Oscillators. As power amplifiers, klystrons have been used over a range of frequency from 500 mc to higher than 10,000 mc with power outputs ranging up to 5 kw continuous operation and up to ten megawatts pulsed operation.[37] Efficiencies as high as 35–40% have been obtained, and 20–25% is not uncommon in the modern tubes. For amplifiers where high power and/or efficiency is of importance the klystron, at present, seems preeminent. Some work has been done on special types of traveling wave tubes which seems to indicate comparable efficiencies and powers as high as perhaps a kilowatt. This work is still at too early a stage to indicate what may be possible in this direction.

It seems probable that the maximum efficiency quoted above can be pushed slightly higher for klystrons. Most of the work on high power C.W. or pulsed amplifiers has been done since the end of the war. The efficiencies and powers quoted were obtained by cascade bunching: i.e., two bunchers, gridless gaps, and well-focused beams. In these more recent tubes the operating conditions have begun to approach the ideal conditions upon which the theoretical analysis is based and their efficiencies approach the theoretical values quite closely. In the light of this progress and the verification of the theory it seems possible to indicate what future trends in frequency and power may be. First, as to frequency, in all microwave tubes the limitation on frequency arises from shrinking of dimensions as the frequency increases. To keep transit angles small, the dimensions must be small, with the ratio of linear dimensions to operating wavelength being proportional to the square root of the operating voltage. In any such device, for a given voltage there will be some minimum current required for successful operation and therefore some minimum input d-c power. If the geometry is such that some or all of the unconverted d-c energy must be dissipated by the radio-frequency structure there is a possibility that the structure cannot dissipate this energy. For a gridless klystron amplifier, by a combination of electrostatic and magnetic focusing, it is possible in principle to get zero interception of the beam by the radio-frequency structure. In practice one can come fairly close to this ideal. On this assumption as to practical focusing, it is possible to use large d-c inputs at relatively high frequencies even though the radio-frequency structures are small. Putting in reasonable numbers for the amount of convergence that can be produced in a well-focused beam and the current densities that can be obtained from available cathode materials, it seems that klystrons will still have considerable gain and efficiency at 30,000 mc in pulsed operation. It is impossible to calculate the high-frequency limit with any precision without making a more careful analysis of the beam focusing problem than has yet been done. It must be stressed again that these predictions depend upon being able to get a high voltage, high density beam completely through the radio-frequency part of the klystron with little or no beam interception. The practicality of this assumption is based on the success that has been attained in the focusing of such beams.

Second, as to power. The ultimate power that may be obtained from a klystron is related to the previous problem of the maximum possible frequency. Here also, because of the geometry of the klystron it seems possible to use very large d-c inputs, and then by good focusing to have the beam dissipate itself in a separate, adequately cooled, and sufficiently large collector with little, or no interception by the radio-frequency structure. Using practical current densities and known data as to focusing possibilities, it seems possible to have amplifiers with input of the order of 100 megawatts for pulsed operation in the wavelength range above 10 cm and several hundred kilowatts input for C.W. operation with efficiencies of perhaps 20 40%. These numbers are not the ultimate possible, however. As was described in a previous section, it is possible to construct klystrons with cavities of the conventional cross section but which extend linearly perpendicular to this cross section. Such cavities are essentially sections of reentrant wave guide. The cathodes used would extend parallel to these cavities, being segments of cylinders producing a line focused beam rather than spherical segments producing a point focused beam. With this geometry the cavities are resonant essentially at the cutoff wave length of the reentrant waveguide. The gap field for this type of resonance is everywhere in phase. With these extended resonators and cathodes it is obviously possible to increase the total beam current by large factors. This current, in fact, is proportional to the length of cathode. Essentially, using this structure is equivalent to putting a large number of conventional klystrons in parallel, inside the same vacuum envelope, and getting the sum of their outputs.

Presumably the only limitation on the length of such a tube would be the probable difficulty of maintaining the cross sections of the cavities sufficiently uniform so that the fields would stay in phase over the whole length. Certainly for lengths of the order of a wavelength or two corresponding to perhaps 5–10 conventional klystrons in parallel, this should not be a real difficulty. By this means one should easily be able to increase the available power by factors of 5 to 10. One can use this sort of structure for multi-cavity oscillators; however, in such cases, if the cavities are too long there is the problem of extra modes which may be excited. The power attained could still be made to reach perhaps 5–10 times that of a conventional structure without running into the extra mode problem.

It seems likely that the future endeavors in klystron research will be in the directions indicated—toward higher powers and high frequencies —both of which seem quite feasible. In addition, in the light of new knowledge about shot noise at microwave frequencies, it appears that a radio-frequency amplifier with a noise figure of 10–12 db is possible. This is another avenue of investigation which would be worth following.

REFERENCES

1. Hamilton, D. R., Knipp, J. K., and Kuper, J. B. H. Klystrons and Microwave Triodes (M.I.T. Radiation Laboratory Series). McGraw-Hill, New York, 1948.
2. Harrison, A. E. Klystron Tubes. McGraw-Hill, New York, 1947.
3. Beck, A. H. W. Velocity-Modulated Thermionic Tubes. Cambridge University Press, Cambridge, England, 1948.
4. Hansen, W. W., Notes on Lectures (given at M.I.T., 1941–44 inc.).
4a. Warnecke, R. R., and Guneard, P. R. Les tubes électroniques à commande par modulation de vitesse. Gauthier Villars, Paris (in press).
5. Hansen, W. W., A type of electrical resonator. *J. Applied Phys.*, **9,** 654 (1938).
6. Hansen, W. W., On the resonant frequency of closed concentric lines. *J. Applied Phys.*, **10,** 38 (1939).

7. Hansen, W. W., and Richtmyer, R. D. On resonators suitable for klystron oscillators. *J. Applied Phys.*, **10,** 189 (1939).
8. Varian, R. H., and Varian, S. F. A high frequency amplifier and oscillator. *J. Applied Phys.*, **10,** 140 (1939).
9. Webster, D. L. Cathode-ray bunching. *J. Applied Phys.*, **10,** 501 (1939).
10. Webster, D. L. The theory of klystron oscillations. *J. Applied Phys.*, **10,** 864 (1939).
11. Hahn, W. C., and Metcalf, G. F. Velocity modulated tubes. *Proc. Inst. Radio Engrs.*, **27,** 106 (1939).
12. Hahn, W. C. Small signal theory of velocity modulated electron beams: wave energy and transconductance of velocity modulated electron beams. *Gen. Elec. Rev.*, **42,** 258, 497 (1939).
13. Alpert, D. Theory of the Monotron. Stanford University, Stanford, California. Unpublished thesis.
14. Arsenjewa-Heil, A., and Heil, O. Electromagnetic oscillators of high intensity *Z. Physik*, **95** (11), 752-762 (1935); (12) (1935).
15. Unpublished reports by Standard Telephone and Cable Laboratories Ltd. 1940-1946.
16. Pierce, J. R. Rectilinear electron flow in beams. *J. Applied Phys.*, **11,** 548 (1940).

    Pierce, J. R. Theory and Design of Electron Beams. Van Nostrand, 1949.
17. Helm, R., Spangenberg, K. R., and Field, L. M. Cathode design procedure for electron beam tubes. *Elec. Commun.*, **24,** 101 (1947).
18. Ginzton, E. L., and Chodorow, M. High Power Pulsed Klystron Project, Report No. 3, ONR Contract N6-onr-251, Task IX, May, 1948.
19. Wang, C. C. Electron beams in axial symmetric magnetic and electric fields. Paper presented at the I.R.E. National Convention, New York City, March, 1949.
20. Fremlin, J. H., Gent, A. W., Petrie, D. P. R., Wallis, P. J., and Tomlin, S. G. Principles of velocity modulation. *J. Instn. Elec. Engrs.*, **93,** 875 (1946).
21. Guénard, P. R. Effet de lentille des champs alternatifs dans les tubes à modulation de vitesse. *Ann. Radioélectricité*, **1,** 319 (1946).
22. Guénard, P. R. Echange d'energie entre un faisceau electronique et un champ éléctromagnétique de faible intensité. *Compt. rend.*, **224,** 898 (1947).
23. Feenberg, E. Notes on Velocity Modulation, Sperry Gyroscope Company, Report No. 5221-1093, 1945.
24. Ramo, S. The electronic-wave theory of velocity modulated tubes. *Proc. Inst. Radio Engrs.*, **27** (12), 757 (1939).
25. Feenberg, E., and Feldman, D. Theory of small signal bunching in a parallel electron beam of rectangular cross section. *J. Applied Phys.*, **17,** 1025 (1946).
26. Warnecke, R. R., Bernier, J., and Guénard, P. R. Groupement et degroupement au sein d'un faisceau cathodique injecté dans un espace exempt de champs extérieurs après avoir été modulé dans sa vitesse I. *J. phys., radium*, **4,** 5, 96 (1943).
27. Guénard, P. R. Sur la possibilité d'une focalisation purement électrostatique dans un tube à modulation de vitesse à conversion par glissement, *Ann. Radioélectricité*, **1,** 74 (1945).
28. Warnecke, R. R., Guénard, P. R., and Fauve, C. Sur les effects de la charge d'espace dans les tubes à modulation de vitesse à groupement par glissement, *Ann. Radioélectricité*, **2,** 224 (1947).

29. Warnecke, R. R., and Bernier, J. Contribution à la théorie des tubes à commande par modulation de vitesse et autres tubes à temps de transit. *Rev. gén. élec.*, **51,** 43, 117 (1942).
30. Warnecke, R. R., Guénard, P. R., and Fauve, C. Sur le rendement des tubes à modulation de vitesse. *Ann. Radioélectricité*, **14,** 303 (1948).
31. Warnecke, R. R., and Guénard, P. R. Sur l'aide que peuvent apporter en télévision quelques récentes conceptions concernant les tubes électroniques pour ultra hautes fréquences. *Ann. Radioélectricité*, **3,** 259 (1948).
32. Pierce, J. R. Reflex oscillators. *Proc. Inst. Radio Engrs.*, **33,** 112 (1945).
33. Ginzton, E. L., and Harrison, A. E. Reflex-klystron oscillators, *Proc. Inst. Radio Engrs.*, **34** (1946).
34. Pierce, J. R., and Shepherd, W. G. Reflex oscillators, *Bell System Tech. J.*, **26,** 460 (1947).
35. Bernier, J. Sur le rendement de conversion des tubes à modulation de vitesse du type reflex. *Ann. Radioélectricité*, **1,** 359 (1946).
36 Chodorow, M. Theory of Reflex Tubes, Part I, Sperry Gyroscope Company Report No 5221 1079, 1946.
37 Chodorow, M , Ginzton, E L., Neilsen, I., Sonkin, S. High Power Pulsed Klystron, I.R E. National Convention, March, 1950

Electronic Theory of the Plane Magnetron*

L. BRILLOUIN

International Business Machines Corporation, New York

CONTENTS

Page
I Introduction 85
II Steady Case 89
III Statement of the Problem. A Method of Integration Similar to Llewellyn's Procedure for a Diode 91
IV Discussion of the Results. Standard Static Characteristics 92
V Double Stream Solutions 97
VI Transients and Oscillations Keeping the Plane Symmetry. Principle of the Method 99
VII Operation of a Magnetron with a Short Impulse of Current 104
VIII. Discussion of British Reports on Similar Problems 108
IX A General Discussion of Electron Trajectories in a Plane Magnetron 114
X Steady Problem: Negative Resistance for Very Low Frequencies 116
XI Small Oscillations of High Frequency: Fundamental Equation for the Trajectories 117
XII Characteristic Impedance of the Oscillating Plane Magnetron 119
XIII Magnetron Impedance for Low Frequencies 121
XIV Magnetron Impedance for High Frequencies 122
XV Discussion of Some Special Examples 124
XVI. Double Stream Electronic Motions. General Formulas 128
XVII Large Resonant Oscillations with Moderate Direct Current 132
XVIII Efficiency and Negative Resistance in One-Anode Magnetrons 135
XIX Physical Meaning of Conditions for Negative Resistance 139
Appendix 142
References 144

I. INTRODUCTION

A typical magnetron structure comprises a cylindrical cathode surrounded by a cylindrical anode, both of finite length. An almost uniform magnetic field parallel to the axis of the tube is provided. When the theory of such a tube is developed, a first approximation is usually made, neglecting the effect of finite length and discussing a structure of infinite length. This is known to represent a crude simplification since the perturbations due to both ends of the tube are fairly large, and it has been

---

* This paper is based on work done for the Office of Scientific Research and Development under contract OEMsr-1007 with Columbia University (1944).

known experimentally that the detailed design of these ends is of importance. A great deal of attention has been devoted to the role of the cylindrical structure and of the ratio of anode to cathode radius. Despite many interesting theoretical facts discovered in this connection, it does not seem that this point is of vital importance, and the disturbance due to the difference in anode and cathode dimension is often of the same order of magnitude as the perturbation due to end effects.

Most modern magnetrons are built with a small anode to cathode distance, the ratio of the radii being less than 2, and this leads to the conclusion that the main theoretical results could be obtained from a discussion of the plane case, when the anode to cathode distance is considered as very small when compared with the average radius.

A similar situation was found in the problem of other electronic tubes (diode, triode, tetrode) where a discussion of the plane problem proved very instructive and lead to interesting results. The plane structure offers the great advantage of yielding much simpler equations, which in many cases can be solved rigorously. Thus the physical difficulties of the problem, which are serious, do not mix with mathematical complexities, and one can avoid a superposition of approximations, which very often proved misleading, or at least very confusing.

The problem is always to obtain a "self-consistent" field distribution, namely one which produces electron trajectories from which a space charge distribution is obtained, that again produces the assumed field distribution. Field equations are usually hard to discuss, but the equations for "self-consistent trajectories" are not too difficult to work with and can be solved rigorously if the current is chosen as a primary datum. This is very similar to the situation found by Llewellyn for other electron tubes, where he was able to obtain a solution for the anode voltage as a function of the current, while the inverse problem of finding the current corresponding to a given voltage was impractical.

The solution of the steady case (constant current and voltage) has been known for some time and is summarized in Section II. Two different types of solutions are found when the voltage is below cutoff:

A. A "double stream" solution, for which there is a constant flow of electrons from the cathode and an equal flow of electrons back to the cathode, resulting in a total current naught.

B. A "single stream" solution, with no electrons emitted from the cathode, but a constant flow of previously emitted electrons, which run parallel to cathode and anode, never reaching any of these electrodes.

Solution A was first proposed by Hull and has been very carefully discussed by Hartree and Stoner in the first British reports of the Series C.V.D. Mag.[9]

Solution B was first found by Hull (1924) and its importance was emphasized by the author of this report in a series of papers.[5]

The question arises of which type is stable or, better said, which one is actually obtained when the magnetron is lit up and put into operation. This problem is discussed in Sections VI, VII, VIII. The theory proves that the space charge will, after some time, settle down on a "single stream" motion of type B, and there does not seem to be any practical way of generating a "double stream" motion A below cutoff.

A general discussion is given in Section IX of conditions where electron trajectories may or may not cross each other. There are some difficulties with problems where negative currents may flow through the tube during certain time intervals. If *no negative currents* are allowed, the following *theorem* is proved.

For single anode plane magnetrons operated *below saturation* with space charge limited currents, the single stream B solution is the only possible one, and *electron trajectories never cross each other.* If saturation is reached and the current is temperature limited, trajectories will intercross and some sort of double stream motion will appear. This is true for any arbitrary law of current in a single anode structure. Multianode structures would behave differently.

The steady characteristic (Fig. 3, Sections IV and X) exhibits some regions with negative resistance, thus showing a possibility to use a single-anode magnetron for sustaining oscillations in an outer circuit. Steady characteristics can be used only for a discussion of low-frequency operation.

The general method of integration developed in Section VI is used in Section XI for the solution of a problem where small oscillations of arbitrary frequency are superimposed upon given static conditions. This enables one to compute the characteristic impedance of a plane one-anode magnetron for small oscillations (Section XII). The case of a low frequency (Section XIII) shows complete agreement with the results obtained in Section X from the static characteristic. For high frequencies (Section XIV) a general formula is obtained, and conditions leading to negative resistance are discussed.

These general formulas have been used for the discussion of several special cases, and the corresponding types of trajectories are shown in Sections XV and XVII while Section XVI is devoted to the problem of double stream solutions that appear in some of the examples chosen for discussion.

These special cases justify a statement made from the general theory that negative resistance cannot be obtained at exact resonance; a small difference between the proper frequency of the magnetron and the frequency of oscillation is needed and the beats between free and forced oscillations play a very important role in conditions for negative resistance.

A comparison of these rigorous results with those previously obtained by different authors is very instructive. It definitely shows what was wrong in the approximations used by many scientists (including the present writer). The point is worth explaining since similar difficulties certainly occur in the more complicated (and much more important) problem of multianode magnetrons. The magnetron structures exhibit some internal resonance frequencies and the plane problem is simpler than the cylindrical case, since it has only one frequency which is just twice the Larmor frequency. When a perturbation is applied, as for instance a periodic perturbation of a given frequency ω, an equation is found which is similar to that of an harmonic oscillator with right-hand term

$$y + 4\omega_H^2 y = f(t)$$

$$\omega_H = \frac{1}{2}\mu_0 \frac{e}{m} H \text{ Larmor frequency}$$

where μ_0 = permeability,
e = electric charge
m = mass of the electron
H = magnetic field

There is no damping term (sy) in the equation, but this is due to some oversimplifications (infinite plane structure), and in practical problems one may assume a certain amount of damping caused by radiation through both ends of the magnetron or by resistances in the outer circuits.

When dealing with such an equation the mathematician builds the solution with a forced oscillation (satisfying the equation with right-hand term) plus a free oscillation, adjusted to fit the initial condition. Engineers call these terms "permanent" and "transient," since they always have to do with damped oscillators.

In the magnetron problem, it was generally assumed that the transient could be safely ignored, since there must be some damping in the actual problem, and it seemed to be only a matter of waiting long enough to give the transient time to die out. The underlying assumption was that the delay could be counted as the time elapsed since the tube was put into operation.

But actually the time delay which comes in is not the time of operation, but the transit time for electrons between cathode to anode—and this transit time is not so very large. This makes it necessary to keep *both transient and permanent* terms in the solution.

Such circumstances could hardly be revealed by a discussion of systems which had been submitted to approximations of different kinds from the beginning. The great advantage of the plane problem is that it can be discussed rigorously throughout, and in this case there is no doubt about the right way to build the actual solution.

From this experience one may conclude that all treatments of the multianode magnetron based on perturbation methods should be taken with great caution, while discussions based on actual computation of trajectories (Hartree, Stoner) should be much more reliable, even when the computation is not very accurate in itself The most important feature is to satisfy exactly the conditions on the cathode.

II STEADY CASE

The theory of the plane magnetron with space charge represents a very interesting special case of the general problem of magnetrons, since it is possible to find a complete rigorous solution involving no approximation. The solution for the steady case was apparently first discovered by Moullin,[1] independently by E. U. Condon and J. C. Slater,[2] then again by Page and Adams.[3]

The present writer found it possible to extend to the case of the plane magnetron Llewellyn's method of integration, which was originally designed for the plane diode without magnetic field.[4]

Let us first summarize the results for the steady case. The current value entering the equations is the sum of the absolute values of two currents which may eventually flow in opposite directions, since equations only yield the absolute value of the velocity component perpendicular to the plane electrodes, but do not say anything about its sign. The general situation in that respect is very similar to the one encountered by Fay, Samuel, and Shockley in their paper on the diode.[4]

The different possibilities are:

(1) Electrons move continuously from cathode to anode, thus yielding a direct current on the anode.

(2) Electrons start from the cathode, reach a virtual cathode, and turn back to the cathode.

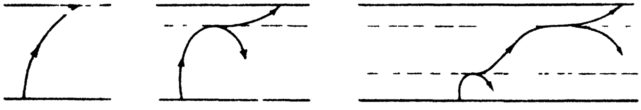
(3) In case many virtual cathodes are found between cathode and anode, electrons always make a stop on each virtual cathode, and may then start either forward or backward. Hence each virtual cathode offers a possibility for partial transmission and partial reflection of the electrons. Loops are impossible, since the x component of velocity cannot be reversed. Only the y component may change sign. On the whole,

the magnetron exhibits successive regions of bunching and debunching of the electron beam.

The total current density I flowing through a unit square area of the tube is a constant. It may be obtained as the superposition of two current densities I_1 and I_2 flowing in opposite directions.

$$\begin{aligned} I &= I_1 - I_2 \\ J &= I_1 + I_2 \geq |I| \end{aligned} \tag{1}$$

The current density entering the equations is the sum J of the absolute values, and not the total current density I. Between two successive virtual cathodes, I_1 and I_2 are separately constant, and so is J. On both sides of a virtual cathode, I obtains the same value, but J may make a sudden change. J rules the whole space charge and voltage distribution (see Fig. 1*c*). The problem of *cutoff conditions* is very interesting to consider. Two different types of solutions have been proposed.



a
no virtual cathode

b
one virtual cathode

c
two virtual cathodes

FIG. 1

A. *Double stream* motion, where electrons leave the cathode, reach a virtual cathode somewhere, in front of the anode, and turn back to the cathode. The total current I is zero and

$$J = 2I_1 \qquad I_1 = I_2 \tag{2}$$

Inside the space charge, half of the electrons are moving up and half are moving down.

$$I_1 = \tfrac{1}{2}\rho\dot{y} \tag{3}$$

where $\pm\dot{y}$ are the velocities of both groups of electrons.

B. *Single stream* motion, where electrons move along straight lines parallel to cathode and anode.

$$I_1 = I_2 = 0, \quad \dot{y} = 0, \quad J = 0 \tag{4}$$

There is no electronic emission from the cathode and no back bombardment either. This solution was proposed by the present writer in a series of papers.[5]

The B solution appears as the natural limit of single stream motions (Fig. 1) with final anode current, when the magnetic field is increased and the current progressively drops to zero.

$$J = I_1 = I \neq 0, \quad I_2 = 0 \tag{5}$$

III. STATEMENT OF THE PROBLEM: A METHOD OF INTEGRATION SIMILAR TO LLEWELLYN'S PROCEDURE FOR A DIODE

The magnetron consists in a plane cathode ($y = 0$) and a plane anode ($y = d$), with a magnetic field H in the z direction. Electrons move in the xy plane since the electric field, in the steady case, is always along y and the Lorentz force has an x component. The x motion can be readily integrated:

$$m\ddot{x} = \mu_0 e H \dot{y}, \quad e < 0 \quad \text{MKS units} \tag{6}$$

$$\dot{x} = \mu_0 \frac{e}{m} H y = -2\omega_H y, \quad (\omega_H = \text{Larmor's angular velocity}) \tag{7}$$

$$\omega_H = -\frac{1}{2}\mu_0 \frac{e}{m} H \tag{8}$$

There is no integration constant in eq. 7 since electrons are emitted on the cathode $y = 0$ without velocity.

Calling $V(y)$ the electric potential, and taking $V = 0$ on the cathode, we may now write the energy relation

$$\dot{x}^2 + \dot{y}^2 = 4\omega_H^2 y^2 + \dot{y}^2 = -\frac{2e}{m} V \tag{9}$$

Let us first assume a single flow of electrons, with $y \geq 0$ and no electrons flowing backwards toward the cathode (B motion, eq. 5).

The whole problem can be discussed along a line familiar to electronics engineers if one proceeds from the well-known Llewellyn formula:

$$\frac{dE}{dt} = \frac{dE}{dy}\dot{y} = \frac{\rho}{\epsilon_0} v = -\frac{J}{\epsilon_0}, \quad E, \text{ electric field} \tag{10}$$

where d/dt is a derivative taken along the path of an electron. Hence, for a constant J value,

$$E = -\frac{J}{\epsilon_0}\tau, \quad \tau, \text{ transit time} \tag{11}$$

and the motion of the electron in the y direction, including the Lorentz force, is

$$\ddot{y} = 2\omega_H \dot{x} + \frac{e}{m} E = -4\omega_H^2 y - \frac{e}{m}\frac{J}{\epsilon_0}\tau \tag{12}$$

using eq. 7. This is the equation of a harmonic oscillator of frequency $2\omega_H$ acted upon by an external force proportional to the time τ. The solution must start from $y = 0$, $\dot{y} = 0$, $\ddot{y} = 0$, for $\tau = 0$ (no velocity, no acceleration on the cathode), hence:

$$y = -\frac{e}{m}\frac{J}{\epsilon_0}\frac{1}{8\omega_H{}^3}[\psi - \sin\psi], \quad \psi = 2\omega_H\tau \tag{13}$$

Once y is obtained, x can be found by integration of eq. 7 and the trajectory is defined by the parametric equations:

$$\begin{aligned} \xi &= -\frac{m}{e}8\epsilon_0\omega_H{}^3\frac{x - x_0}{J} = 1 - \frac{\psi^2}{2} - \cos\psi \\ \dot{\xi} &= -2\omega_H\eta \\ \eta &= -\frac{m}{e}8\epsilon_0\omega_H{}^3\frac{y}{J} = \psi - \sin\psi \end{aligned} \tag{14}$$

This represents the motion of an electron always traveling forward (single stream B motion).

The possibility of other solutions (double stream A motion) where some electrons flow back toward the cathode is discussed in another section. It involves a different definition of the current J in eqs. 10 and 11 and the use of a double sign in eq. 13:

$$\psi = \pm 2\omega_H\tau \tag{15}$$

which yields a double sign for $\dot{y}$.

One should notice that any reversal of the motion must take place on the *virtual cathodes*, when the velocity $\dot{y}$ and the acceleration $\ddot{y}$ are both zero:

Virtual cathodes P (eventual motion reversal)

$$\psi = 2n\pi = \eta, \quad \eta = 0, \quad \dot{\eta} = 0, \quad \dot{\xi} = -4n\pi\omega_H \tag{16}$$

The $\dot{\eta}$ velocity is maximum on the following layers, which we shall call anticathodes Q.

$$\psi = (2n + 1)\pi = \eta, \quad \dot{\eta}_{\max} = 4\omega_H, \quad \ddot{\eta} = 0, \quad \dot{\xi} = -(4n + 2)\pi\omega_H \tag{17}$$

IV. DISCUSSION OF THE RESULTS: STANDARD STATIC CHARACTERISTIC

Equations 14 represent a cycloid and its image (± sign) (Fig. 2*a*), and the motion of electrons corresponds to a uniform generation of the cycloid. The cathode is at the origin ($\psi = 0$), and similar conditions are found again on the virtual cathodes eq. 16, but for the fact that the x component of the velocity does not vanish on these virtual cathodes.

These are the only positions where electrons may actually reverse their motion, with a change in the ± sign of the component $\dot{y}$ of the velocity. These circumstances explain and justify our assumptions of Section II. Figure 2*b* shows y as a function of ψ, which is proportional to the transit time.

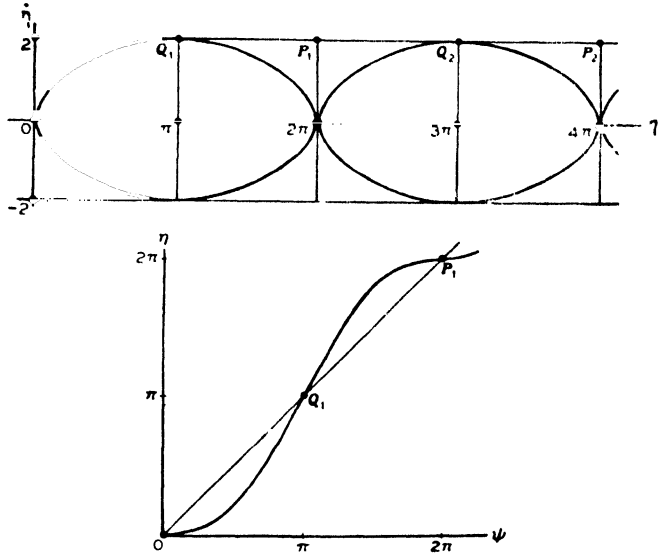


FIG. 2

We must now return to the electric potential V itself (Eq. 9).

$$-\frac{2e}{m}V = 4\omega_H^2 y^2 + \dot{y}^2 - 4\omega_H^2 y^2 + \left(\frac{eJ}{4\epsilon_0 m\omega_H^2}\right)^2 (1 - \cos\psi)^2$$

$$= \left(\frac{eJ}{4\epsilon_0 m\omega_H^2}\right)^2 [(\psi - \sin\psi)^2 + (1 - \cos\psi)^2] \qquad (18)$$

Let us take $y = d$ and discuss the relation between the voltage V and the current J on the anode for a given magnetic field (given ω_H). For low voltages we obtain no current

$$-\frac{2e}{m}V \le 4\omega_H^2 d^2, \quad J = 0, \quad V_{co} = -\frac{2m\omega_H^2}{e} d^2 \qquad (19)$$

These are the "cutoff" conditions. For higher voltages we obtain a current J from eq. 18, where ψ must be computed from conditions (13) and (14). A small current J means a very large ψ, then ψ decreases when J increases, and ψ goes down to zero for very large current values.

Accordingly, $(-2e/m)V$ oscillates between the limits given by $\psi = 2n\pi$ (*virtual cathodes*) or $\psi = (2n + 1)\pi$ (which were called *anticathodes*)

$$4\omega_H{}^2d^2 \leq -\frac{2e}{m}V \leq 4\omega_H{}^2d^2 + \frac{e^2J^2}{4\epsilon_0{}^2m^2\omega_H{}^4} \tag{20}$$

For very large currents, small ψ values are reached and power expansions of $\sin\psi$ or $\cos\psi$ can be used in eqs. 14 and 18. The result is

$$V = \frac{3}{4}\left(\frac{6m}{e}\right)^{1/3} d^{4/3}\left(\frac{J}{\epsilon_0}\right)^{2/3} + \frac{1}{10}V_{co} \tag{21}$$

where V_{co} is the cutoff value (19). The first term in eq. 21 is Langmuir's voltage for a plane diode without magnetic field

The characteristic curve shown on Fig. 3 summarizes the results. It has been drawn with dimensionless variables by a convenient choice of units:

Unit of voltage: the cutoff voltage V_{co} (eq. 19)
Unit of current: the current obtained, at V_{co} in a plane diode without magnetic field, according to Langmuir (22)

Hence the reduced voltage is given by

$$U = \frac{V}{V_{co}} = -\frac{e}{m}\,\frac{V}{2\omega_H{}^2d^2} \tag{23}$$

The reduced current is represented by

$$C = -\frac{e}{m\epsilon_0}\,\frac{9J}{16\omega_H{}^3d} \tag{24}$$

while the final asymptotic curve, eq 21, for large currents now reads

$$U = C^{2/3} + \tfrac{1}{10} \tag{25}$$

The standard characteristic curve of Fig. 3 shows the $P_1P_2P_3$ points (*virtual cathodes*, eqs. 16 and 20) and the $Q_1Q_2Q_3$ points corresponding to *anticathodes*, eqs. 17 and 20. It is a typical S-shaped curve with *negative resistance* along P_1Q_2 or P_2Q_3.

If the voltage is increased above cutoff and no oscillations can possibly start, the current should suddenly increase to its P_1 value, $C = 0.72$. The parametric equation of the whole curve is obtained when C and U are expressed as functions of ψ. Equation 14 yields

$$C = \frac{4.5}{\psi - \sin\psi} = \frac{4.5}{\delta} \tag{26}$$

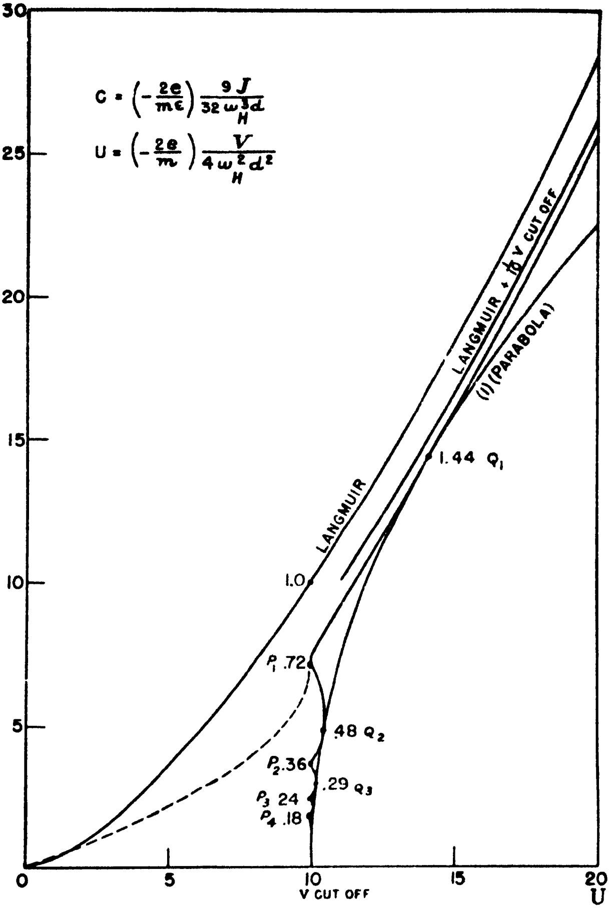


FIG. 3

where δ is the reduced anode distance according to eq. 14

$$\delta = -8\frac{m}{e}\frac{\epsilon_0}{J}\omega_H^3 d = \frac{4.5}{C}$$

Next we obtain from eqs. 18, 23, and 24

$$W = \frac{81}{4}\frac{U}{C^2} = -32\frac{m}{e}\left(\frac{\epsilon_0}{J}\right)^2\omega_H^4 V = (\psi - \sin\psi)^2 + (1 - \cos\psi)^2 \quad (27)$$

and

$$U = 1 + \left(\frac{1 - \cos\psi}{\psi - \sin\psi}\right)^2 \quad (28)$$

A plot of W versus reduced distance δ is shown on Fig. 4 and gives the voltage distribution inside the space charge for a single stream B motion. The curve always climbs up with positive curvature and oscillates between the P parabola ($W = \delta^2$) and the Q parabola ($W = \delta^2 + 4$) according to eq. 20.

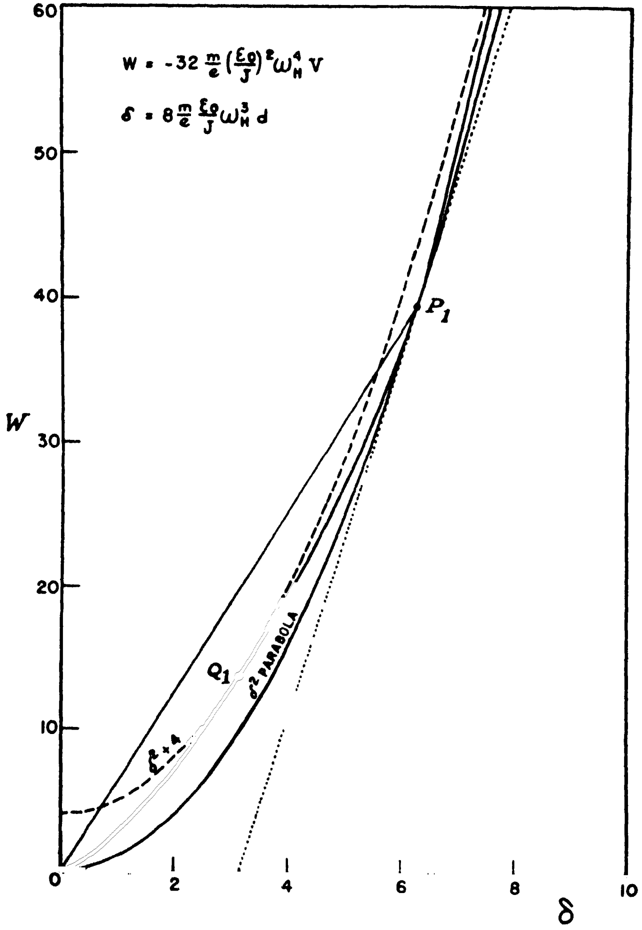


Fig. 4

Virtual cathodes P correspond to

$$\begin{array}{llllll} \psi = 2\pi & 4\pi & 6\pi & \cdots & 2n\pi & \\ C = 0.718 & 0.358 & 0.239 & \cdots & \dfrac{0.718}{n} & \\ U = 1 & & & & & \end{array} \tag{29}$$

Anticathodes Q are found for

$$\begin{array}{llllll} \psi = \pi & 3\pi & 5\pi & \cdots & (2n+1)\pi & \\ C = 1.436 & 0.48 & 0.288 & \cdots & \dfrac{1.436}{2n+1} & \\ U = 1.405 & 1.045 & 1.0016 & \cdots & 1 + \dfrac{0.405}{(2n+1)^2} & \end{array} \tag{30}$$

The discussion recently given by Page and Adams[3] refers to the case when there is only one virtual cathode.

V. Double Stream Solutions

In case of a double stream motion, type A, Section II, we must use the definition of eq. 1 and divide the space charge into two parts.

$$\rho = \rho_1 + \rho_2 \tag{31}$$

ρ_1 corresponds to electrons with a positive velocity $\dot{y}_1$ and ρ_2 for negative velocity $\dot{y}_2$. In order to obtain a unique solution for the potential V (eqs. 9 and 18), we have

$$\dot{y}_2 = -\dot{y}_1 \tag{32}$$

hence

$$I_1 = \rho_1\dot{y}_1 \qquad I_2 = \rho_2\dot{y}_1$$

$$I = I_1 - I_2 = (\rho_1 - \rho_2)\dot{y}_1 \quad \text{total current} \tag{33}$$

$$J = I_1 + I_2 = (\rho_1 + \rho_2)\dot{y}_1 = \rho\dot{y}_1 \tag{34}$$

Llewellyn's formula (10) can still be used, provided the J value is substituted for the I total current. Since we already wrote J, eqs. 10 to 14 are unchanged, but for the introduction of a double sign in eq. 15.

We may now come back to the discussion of *cutoff* conditions A zero total current may be obtained, as stated in Section II, Eqs. 2 and 5, in two different ways:

A with $J \neq 0$, $\quad \dot{y}_1 \neq 0$, $\quad I_1 = I_2 = \frac{1}{2}J$
B with $J = 0$, $\quad \dot{y}_1 = 0$

In case B, the V,y curve is the lower parabola of Fig. 1, and this solution corresponds to the one given by the present writer in a former paper.[5] In case A, the V,y curve is the wavy curve $0Q_1P_1Q_2, \cdots$ of Fig. 4, where the anode must be located at one of the virtual cathodes P_1P_2, $\cdots$, in order that electrons just reverse their motions in front of the anode. But these points, $P_1P_2, \cdots$, correspond to the beginning of the instability curves of Fig. 3 with negative resistance. Hence these positions are unstable and the stable conditions should be represented by case B. Electrons are moving on straight lines parallel to cathode and anode

$$\begin{aligned} y &= \text{Constant} \\ \dot{x} &= -2\omega_H y \text{ (Eq. 7)} \end{aligned} \tag{35}$$

and there is a uniform space charge density

$$\begin{aligned} \rho_0 &= \epsilon_0 \frac{m}{e} 4\omega_H{}^2 \\ V &= -\frac{m}{e} 2\omega_H{}^2 y^2 \end{aligned} \tag{36}$$

This is a special case of a solution given by the present writer for the cylindrical magnetron [6] where electrons run on circular orbits around the filament, while the space charge exhibits a density ρ_0 near the filament, dropping progressively to $\frac{1}{2}\rho_0$ at large distance from the filament.

One may notice on Fig. 3 an additional dotted curve running from the origin to the first P_1 point (virtual cathode). Similar curves would also run from 0 to P_2, P_3, their meaning results from the following discussion.

Let us suppose the magnetron to be first operated with high voltage and high current. Then by decreasing the voltage down to the cutoff, our representative point comes down along the Q_1P_1 curve and reaches the point P_1 when a virtual cathode is built up just in front of the anode. This corresponds to a certain current $C = 0.72$ on Fig. 3. Here we must remember that C is related with the total absolute current J as defined in eq. 1 or 34. While we were following down the Q_1P_1 branch, there was an electronic current running one way, with electrons flowing from cathode to anode. Now, at P_1, a virtual cathode is obtained, where electrons may revert their course and run back to the cathode. This means that the space charge distribution obtained between cathode and anode may correspond to any one of the two following cases

$$\left.\begin{aligned} &\text{I. One-way current } +J \\ &\text{II. Two-way currents } +\tfrac{1}{2}J \; -\tfrac{1}{2}J \end{aligned}\right\} \quad C = 0.72 \tag{37}$$
$$I = 0$$

This last type of current distribution can be maintained for lower voltages, below cutoff, in which case the space charge will extend from $y = 0$ to $y = y_l$ with a virtual cathode at y_l and vacuum without space charge for $y_l < y < d$. This means the following voltage distribution

$0Q_1P_1$ curve of Fig. 4, $0 < y < y_l$
Straight line (no space charge), $y_l < y < d$

The straight line is the tangent t_0 the parabola at the point P_1 of Fig. 4, in order to insure continuity of V and $\partial V/\partial y$ at the virtual cathode P_1. Its equation reads

$$-\frac{2e}{m} V_{t0} = 8\omega_H^2 y_l \left(y - \frac{y_l}{2}\right) \tag{38}$$

and y_l is obtained from the relation giving a virtual cathode at y_l, point P_1 (eq. 16)

$$-\frac{2e}{m\epsilon_0} J = \frac{8\omega_H^3}{\pi} y_l \tag{39}$$

or, according to eq. 24

$$C = \frac{9}{4\pi}\frac{y_l}{d}$$

We now substitute in eq. 38 and take $y = d$ on the anode, hence

$$U = -\frac{2e}{m} V \frac{1}{4\omega_H^2 d^2} = 2\frac{y_l}{d}\left(1 - \frac{y_l}{2d}\right) = \frac{8\pi C}{9} - \left(\frac{4\pi C}{9}\right)^2 \tag{40}$$

This is the parabola drawn as a dotted curve on Fig. 3. As explained before, it represents the same type of space charge distribution as obtained at P_1, but now realized with currents running both ways, while the space charge shrinks as a whole toward the cathode when the anode voltage is decreased. Whether this space charge and current distribution can really be obtained by simply decreasing the anode voltage below cutoff is an open question.

VI. TRANSIENTS AND OSCILLATIONS KEEPING THE PLANE SYMMETRY: PRINCIPLE OF THE METHOD

The problems discussed in the preceding sections were limited to steady conditions, with constant currents and voltages. It is now necessary to extend the discussion to the case of varying currents and voltages, and we shall first assume that the plane symmetry is maintained, all varying quantities being only functions of y and t, *but not of* x. Under such conditions the equations of motion (7) and (12) are maintained

$$\begin{aligned} \dot{x} &= -2\omega_H y \\ \ddot{y} + 4\omega_H^2 y &= \frac{e}{m} E \end{aligned} \tag{41}$$

We must introduce the displacement current into the definition of current

$$-I = \rho\dot{y} + \epsilon_0 \frac{\partial E}{\partial t} \tag{42}$$

The minus sign yields a positive current for electrons moving toward the anode. We assume a one-way flow of electrons. The Poisson relation

$$\frac{\partial E}{\partial y} = \frac{\rho}{\epsilon_0} \tag{43}$$

together with eq 42 yields the Llewellyn formula

$$\frac{dE}{dt} = \frac{\partial E}{\partial t} + \frac{\partial E}{\partial y} \dot{y} = -\frac{I}{\epsilon_0} \tag{44}$$

As in eq. 10 d/dt is the time derivative taken along the path of an electron. The important point is that the *total current I is a function of t only and does not depend upon y.* This results directly from Maxwell's theory of the displacement current.

Let us use as parameters, the *time t_0* at which *an electron leaves the cathode* and the *transit time τ*. The time t at which this electron is observed is

$$t = t_0 + \tau \tag{45}$$

On the cathode (*no saturation*) the electric field is zero, and eq. 44 is readily integrated

$$E(t_0,t) = -\frac{1}{\epsilon_0}\int_{t_0}^{t} I(t)dt = F(t) - F(t_0) \tag{46}$$

where $I(t) = -\epsilon_0(dF/dt)$

$$F(t) = -\frac{1}{\epsilon_0}\int_0^t I(t)dt$$

This condition applies only inside the space charge. It may happen that the space charge does not fill in completely the cathode-anode interval. In such a case the position $y_f(t)$ of the boundary, at time t, must be computed together with the field $E_f(t)$ on this boundary. In the vacuum, outside the space charge $y > y_f(t)$, a constant field equal to $E_f(t)$ is obtained.

When eq. 46 is introduced in eq. 41, together with the initial conditions

$$y = 0, \quad \dot{y} = 0, \quad \text{at} \quad \tau = 0, \quad t = t_0 \tag{47}$$

meaning no saturation on the cathode, the equation can be integrated and yields $y(t_0,t)$, namely, the position at time t of an electron which left the cathode at t_0. The general solution can be written in the following way: let us call $f(t)$ a function such as

$$\ddot{f} + 4\omega_H{}^2 f = \frac{e}{m} F(t) \tag{48}$$

then our general solution is

$$y(t_0,t) = f(t) - \frac{e}{m4\omega_H^2}F(t_0) + C\cos 2\omega_H(t - t_0) + S\sin 2\omega_H(t - t_0) \quad (49)$$

where C and S must be chosen to fit the initial conditions eq. 47

$$C = -f(t_0) + \frac{e}{m4\omega_H^2}F(t_0), \quad S = -\frac{1}{2\omega_H}\dot{f}(t_0) \quad (50)$$

What is needed next is the potential $V(y,t)$ at a point y and time t. Let us consider another point y_1 at t and call θ the time t_0 at which electrons left the cathode to reach y_1 at t.

$$V(y,t) = -\int_{y_1=0}^{y_1=y} E(y_1,t)dy_1 = -\int_{\theta=t}^{\theta=t_0} E(\theta,t)\left(\frac{\partial y_1}{\partial\theta}\right)_{t\text{ const}} d\theta \quad (51)$$

or, using eq. 46,

$$V(y,t) = \int_{t_0}^{t} [F(t) - F(\theta)]\left(\frac{\partial y_1}{\partial\theta}\right)_{t\text{ const}} d\theta \quad (52)$$

The important point is that we can compute $(\partial y_1/\partial\theta)_{t\text{ const}}$ from eq. 49, in the general case

$$\left(\frac{\partial y}{\partial t_0}\right)_{t\text{ const}} = -\frac{e}{m4\omega_H^2}\dot{F}(t_0) + \cos 2\omega_H(t - t_0)\left(\frac{\partial C}{\partial t_0} - 2\omega_H S\right)$$
$$+ \sin 2\omega_H(t - t_0)\left(\frac{\partial S}{\partial t_0} + 2\omega_H C\right) \quad (53)$$

where

$$\frac{\partial C}{\partial t_0} - 2\omega_H S = \frac{e}{m4\omega_H^2}\dot{F}(t_0) \quad (54)$$

$$\frac{\partial S}{\partial t_0} + 2\omega_H C = -\frac{1}{2\omega_H}\ddot{f}(t_0) - 2\omega_H f(t_0) + \frac{e}{m2\omega_H}F(t_0) = 0 \quad (55)$$

according to eq. 48. Hence, referring to eq. 46

$$\left(\frac{\partial y}{\partial t_0}\right)_{t\text{ const}} = -\frac{eI(t_0)}{m\epsilon_0 4\omega_H^2}(-1 + \cos 2\omega_H(t - t_0)) \quad (e < 0) \quad (56)$$

$(\partial y/\partial t_0)_t$ always keeps the same negative sign, provided the current $I(t_0)$ at the moment of electron emission is positive. Hence, formula 56 proves the following and very important result: If the *current is space charge limited* (eq. 46) and $I(t_0)$ is *positive*, two neighbor *trajectories cannot cross each other.* Electronic layers emitted from the cathode travel without overlapping, a result which justifies our former assumptions after eq. 42. The only possible thing is that *two electron trajectories come into contact,* when

$$\left(\frac{\partial y}{\partial t_0}\right)_t = 0, \quad \cos 2\omega_H(t - t_0) = 1, \quad 2\omega_{H\tau} = 2n\pi \tag{57}$$

This means *infinite space charge density* on certain moving surfaces defined by eq. 57. We shall study this question on some special cases in the next sections. Going back to eq. 52 and using eq. 56 we obtain

$$\left\{\begin{array}{l} V_{(y\,t)} = -\dfrac{e}{m4\omega_H^2}\displaystyle\int_{\theta = t_0}^{\theta = t} [F(t) - F(\theta)][1 - \cos 2\omega_H(t - \theta)]\dot{F}(\theta)d(\theta) \\ I_{(t)} = -\epsilon_0\dot{F}(t), \quad y \text{ corresponds to } t_0,t \end{array}\right. \tag{58}$$

This gives us the potential throughout the space charge. The whole method is a systematic extension of Llewellyn's discussion for a plane diode. This author's results are found when ω_H is taken equal to zero (no magnetic field).

The points to be discussed carefully are:

(1) The position $y_f(t)$ of the *boundary of the space charge* which corresponds to the *first electrons* to leave the cathode, when the total current $I(t_{of})$ first becomes positive.

$$y_f(t_{of},t), \quad I(t_{of}) = 0, \quad \dot{I}(t_{of}) > 0 \tag{59}$$

(2) The trajectories of electrons after the emission has ceased and the *last electrons* remaining in the space charge fall back on the cathode

$$I(t_o) \leq 0 \tag{60}$$

last electrons,

$$t_{ol}, \quad I(t_{ol}) = 0, \quad \dot{I}(t_{ol}) < 0$$

CASE I. (See Fig. 5.) At the *beginning of the discharge* the first electrons emitted are at $y_f(t) = y(t_{of},t)$ between cathode and anode. On this boundary, the electric field is

$$E_f(t) = F(t) - F(t_{of}) \tag{61a}$$

and a constant field E_f extends from the space charge boundary up to the anode ($y = d$). The potential of the anode is

$$V(d,t) = V(y_f,t) - E_f(t)(d - y_f(t)) \tag{61b}$$

During the discharge, when electrons reach the anode, the potential of the anode is given by eq. 58 with $y = d$.

CASE II. At the *end of the discharge*, when emission stops on the cathode and electrons start falling back on it, the limits of integration in eq. 58 must be modified (see Fig. 5). Some electrons, that left the cathode at an earlier time t_{oi} are now just back upon it at time t:

$$t > t_{oi}, \quad y(t_{oi}, t) = 0 \text{ cathode} \tag{62a}$$

and the integral in eq. 58 must be taken from t_{od} to t_{oi} instead of t_{od} to t,

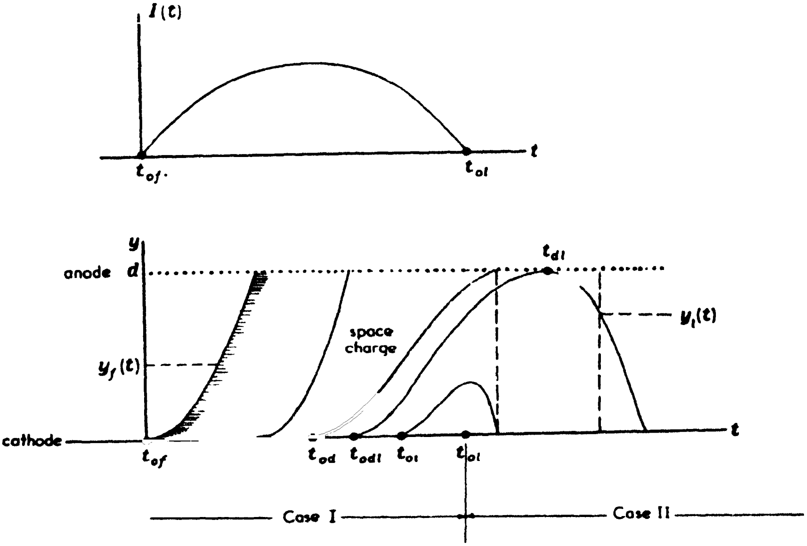


FIG. 5

since the cathode now corresponds to electrons emitted at t_{oi} and no more to electrons emitted at t

$$V(d,t) = -\frac{e}{m4\omega_H^2}\int_{t_{od}}^{t_{oi}} [\cdot\ \cdot\ ][\cdot\ \cdot\ \cdot]\dot{F}(\theta)d\theta \tag{62b}$$

where the brackets are the same as in eq 58

This works during the interval

$$t_{oi} < t < t_{di} \tag{63}$$

if we call t_{di} the time when the last electron reaches the anode. In the case of Fig. 5, it was assumed that $t_{di} > t_{oi}$, electrons being able to reach the anode for some time after the current has been reversed. The opposite situation ($t_{di} < t_{oi}$) may not be excluded.

When t is larger than t_{di}, there is again a moving boundary $y_i(t)$ for the space charge and a vacuum between $y_i(t)$ and anode. Hence the formula for the anode voltage must now read

$$V_{(d,t)} = -\frac{e}{m4\omega_H{}^2}\int_{t_{0dl}}^{t_{0l}} [\cdot\ \cdot\ \cdot][\cdot\ \cdot\ \cdot]\dot{F}(\theta)d\theta - E_l(d - y_l(t)) \tag{64}$$

similar to (61*b*).

Some other troubles might occur in case of overlapping of two successive discharges, if a new discharge would start at a time when some electrons still remain in the tube as residue of former discharges. This will be discussed in special examples.

The whole method is straightforward, provided a certain function $I(t)$ is given for the total current through the tube. This same method was applied by Llewellyn for small oscillations in a plane diode and more recently by Brillouin for large oscillations (class C) in a diode.[7]

As explained before, a certain current $I(t)$ is given, and the space charge and voltage are obtained. The question remains of how to realize $I(t)$ and the corresponding $V(t)$, and how to build the circuits outside of the tube which could yield the necessary relation between I and V. We shall not discuss this problem here.

VII. OPERATION OF A MAGNETRON WITH A SHORT IMPULSE OF CURRENT

There was some uncertainty, in our discussions of Sections IV and V, about the stabler space charge distribution obtained in a magnetron when the anode voltage is below cutoff. Let us discuss this problem in an empirical way. A magnetron is first kept under zero voltage (no space charge). Then a certain current impulse $I(t)$ is impressed on the structure, yielding a finite total charge Q_0. This results in building up a space charge Q_0 and leaving a certain voltage on the anode. We may thus compute the actual structure of the space charge and the final voltage.

The case of an exponential impulse (see Fig. 6)

$$I = \begin{cases} 0 & \text{for } t < 0 \\ I_0 e^{at} & t > 0 \end{cases} \tag{65}$$

was discussed in report.[8] It yields trajectories of the general type sketched on Fig. 7. A final stage is reached after a long time, and exhibits *no current* whatever (no a-c component). Oscillations take place inside the space charge, in such a way that electronic and displacement current cancel each other exactly and *undamped voltage oscillations* may be observed on the anode. The successive sheets of electrons emitted from the cathode never overlap each other.

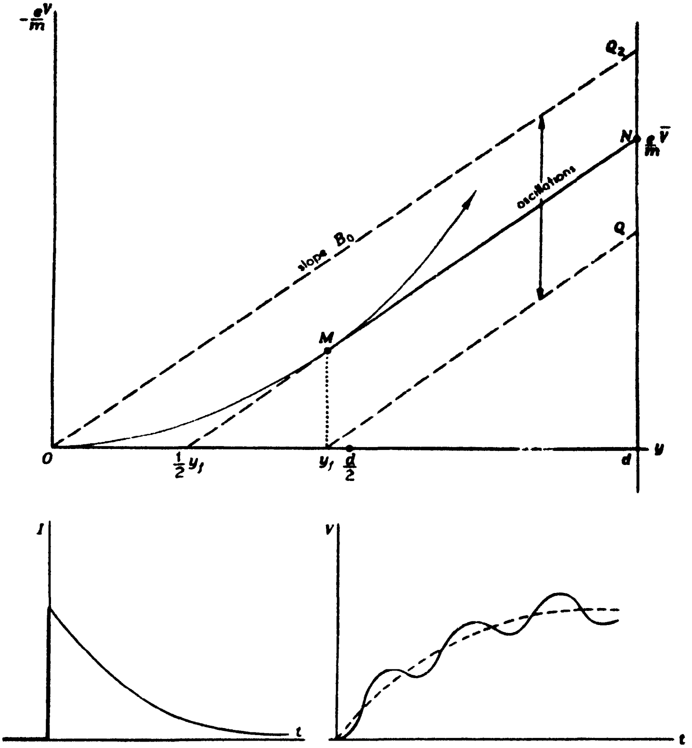

-eV/m
Q₂
N
eV̄/m
oscillations
slope B₀
M
Q
O
½y₁
y₁
d/2
d
y
I
t
V
t

FIG. 6

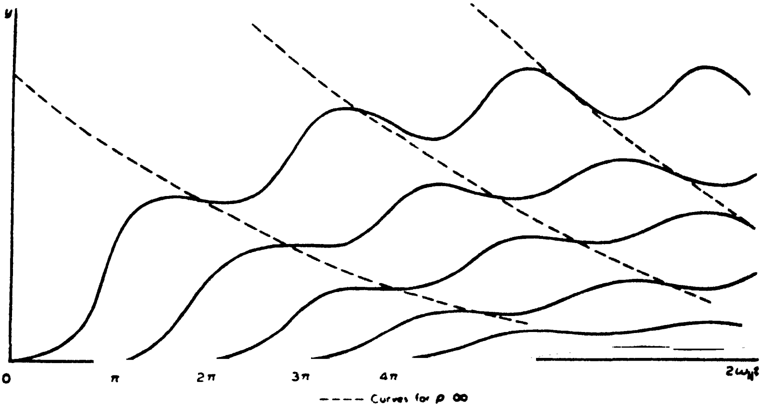

y
0
π
2π
3π
4π
2ω_H t
---- Curves for ρ ∞

FIG. 7

Similar results can be found more easily for a *rectangular current impulse.*

$$I = \begin{cases} 0 & t < 0 \\ I_0 & 0 < t < T \qquad \text{(I)} \\ 0 & T < t \qquad \text{(II)} \end{cases} \tag{66}$$

We apply the general method of Section VI and build the function F of eq. 46

$$F(t) = \begin{cases} 0 & t < 0 \\ -\dfrac{I_0}{\epsilon_0} t & 0 < t < T \\ -\dfrac{I_0}{\epsilon_0} T & T < t \end{cases} \tag{67}$$

hence the $f(t)$ of eq 18

$$f(t) = \begin{cases} 0 \quad t < 0 \quad D = -\dfrac{e}{m} \dfrac{I_0}{4\omega_H{}^2\epsilon_0} \\ Dt \quad 0 < t < T \\ D\left[T + \dfrac{1}{2\omega_H} \sin 2\omega_H(t - T) \right] \qquad T < t \end{cases} \tag{68}$$

The sine term is needed to safeguard the continuity of f and df/dt at T. The solution obtained for $y(t_0, t)$ is zero when $t_0 < 0$ or $t_0 > T$ and obtains a finite value when $0 < t_0 < T$

$$y(t_0, t) = D \begin{cases} \left[t - t_0 - \dfrac{1}{2\omega_H} \sin 2\omega_H(t - t_0) \right] & t_0 < t < T \\ \left[T - t_0 - \dfrac{1}{2\omega_H} \sin 2\omega_H(t - t_0) + \dfrac{1}{2\omega_H} \sin 2\omega_H(t - T) \right] & \\ & T < t \end{cases} \tag{69}$$

These trajectories are seen on Fig. 8 with the coordinates

$$\phi = 2\omega_H t \qquad \eta = \frac{y}{D} 2\omega_H = -\frac{m}{e} \frac{8\omega_H{}^3}{I_0} \epsilon_0 y \tag{70}$$

hence

$$\eta = \begin{cases} \phi - \phi_0 - \sin(\phi - \phi_0) \\ \Phi - \phi_0 - \sin(\phi - \phi_0) + \sin(\phi - \Phi) \end{cases}$$

These are the same dimensionless quantities as used in Section V. In the drawing, it is assumed that $2\omega_H T$ is 3π, giving large oscillations of the space charge boundary. With $\Phi = 2\omega_H T = 2n\pi$ one would obtain a fixed nonoscillating boundary, but the voltage would nevertheless go on oscillating since internal oscillations of the space charge cannot be avoided. Each electronic layer oscillates around the average position.

$$\eta = 2\omega_H(T - t_0) = \Phi - \phi_0 \qquad \Phi = 2\omega_H T \tag{71}$$

Infinite space charge may be observed on the lines

$$\phi - \phi_0 = 2\omega_H(t - t_0) = 2n\pi \tag{72a}$$

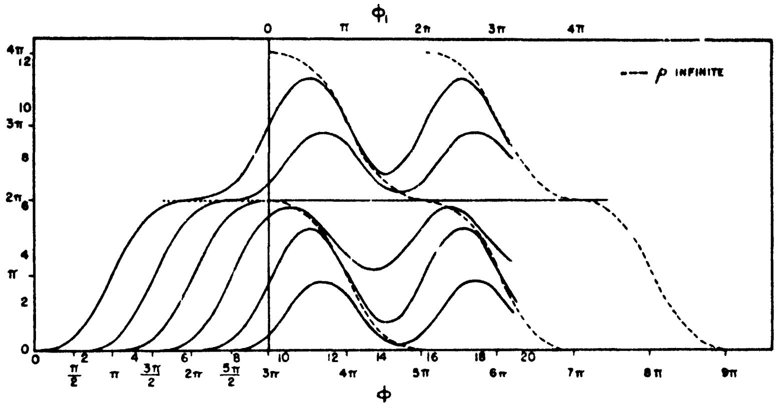


FIG. 8

corresponding to condition (57). This gives

$$\begin{aligned} t < T \quad & \text{horizontal lines} \quad \eta = 2n\pi \\ T < t \quad & \eta = \Phi - \phi + 2n\pi + \sin(\phi - \Phi) \end{aligned} \tag{72b}$$

which gives the curves marked ($\rho \propto$) on Fig. 8.

In all these examples, the same following points are observed: (*a*) oscillating voltage on the anode, in the final stage when the current is identically naught; (*b*) Constant electric field between space charge and anode, hence no displacement current on the anode; (*c*) compensation of electronic and displacement currents inside the space charge; (*d*) no electronic motion on the cathode, hence electronic current zero on the cathode.

The computation of the anode voltage proceeds along the same lines as before. Let us obtain it for the second part of the process, when $t > T$. We must first take eq. 58 which reads

$$V = D\frac{I_0}{\epsilon_0}\int_{t_0}^{T} (T - \theta)(1 - \cos 2\omega_H(t - \theta))d\theta \tag{73}$$

The integral extends from t_0 to T only, since $\dot{F}(\theta)$ becomes zero for $\theta > T$. Taking $t_0 = 0$ we obtain V_f on the space charge border

$$-\frac{e}{m} V_f = D^2 \int_0^\Phi (\Phi - \psi)(1 - \cos(\phi - \psi))d\psi, \quad \psi = 2\omega_H\theta \tag{74}$$

$$-\frac{e}{m} V_f = D^2 \left[\frac{1}{2}\Phi^2 - \Phi \sin \phi + \cos(\phi - \Phi) - \cos \phi\right] \tag{75}$$

and we have to add the integral of the field between y_f and d (eq. 61*b*) in order to obtain the anode potential

$$-\frac{e}{m} V_a = -\frac{e}{m} V_f + \frac{e}{m} E_f(d - y_f) = -\frac{e}{m}\frac{I_0 T}{\epsilon_0} d$$
$$-\frac{e}{m} V_f - D^2\Phi[\Phi - \sin \phi + \sin(\phi - \Phi)]$$

$$-\frac{e}{m} V_a = -\frac{e}{m}\frac{I_0}{\epsilon_0} Td$$
$$+ D^2\left[-\frac{1}{2}\Phi^2 + \cos(\phi - \Phi) - \cos \phi - \Phi \sin(\phi - \Phi)\right]$$

This keeps on oscillating, even in the case $\Phi = 2n\pi$ when the boundary of the space charge is not moving and oscillations are found only inside the space charge. In this example, we thus obtain undamped voltage oscillations with no current oscillations. This means that the resistance of the outside circuit is supposed to be infinite, a condition which does not offer any possibility for energy dissipation. The fact that there is no energy dissipation explains the absence of damping in our example.

The voltage oscillations would die out if the anode were supposed to be connected to the battery through a very high (but not infinite) resistance. After a time the space charge would settle down to the steady motion on parallel layers (B motion, Section V), which actually appears as the stable type.

VIII. DISCUSSION OF BRITISH REPORTS ON SIMILAR PROBLEMS

The question of the space charge distribution obtained in a magnetron operated below cutoff was first discussed in some British reports.[9] As in this paper, the idea was to obtain the space charge distribution realized when the voltage is applied upon the magnetron and a short current impulse brings in the necessary amount of electric charge. These former discussions will be examined now, but it is necessary to stress immediately some differences in the statement of the problem.

(1) The British reports consider either plane or cylindrical magnetrons, with ratios of anode to cathode radius of 2 or 3, while our discussion was limited to the case of a *plane* magnetron (ratio $1 + \epsilon$).

(2) In the British reports, it is assumed that the law of variation of the applied *voltage* is given. Trajectories, space charge, and current are computed by a very tedious process of numerical integration. In our case, we started from a given *current* law and obtained the trajectories, space charge, and voltage.

(3) Since they start from a given voltage law, the British authors assume that the final voltage is a constant. They obtain a final oscillating current, which shows no sign of damping. Hence, the final stage corresponds to a *zero resistance* in the outside circuit. In our problem of Section VII, a final zero current was assumed, yielding an undamped oscillating voltage, which meant an *infinite resistance* in the output circuit.

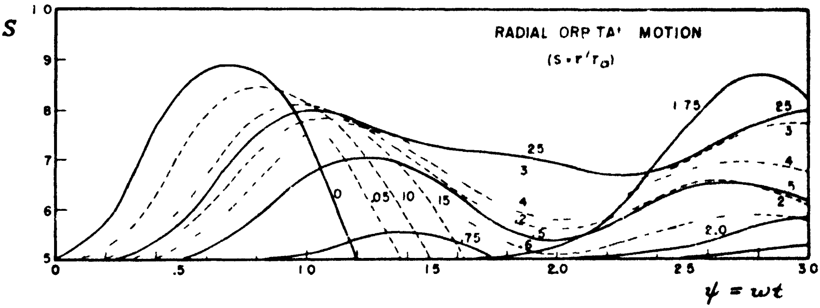


FIG. 9

Both cases exhibit undamped oscillation of some sort, a circumstance related to the fact that there is no possibility for dissipating energy in the circuit outside the magnetron.

In actual conditions of operation, the output circuit would exhibit a finite resistance, hence energy dissipation, and the oscillations inside the space charge would thus be damped and die out (see end of Section VII).

The British authors first discussed the case of a given voltage suddenly applied on the magnetron. In this problem, the current is temperature limited at the beginning, and then becomes gradually space charge limited. For such examples it was found that the electron trajectories would cross each other after a short time. Hartree afterward assumed a voltage gradually applied on the tube in such a way that saturation (temperature-limited) current would never occur, and he obtained noncrossing trajectories.

We may immediately notice the similarity between Hartree's empirical result and our remarks of Section VI, eq. 56. The question of conditions leading to noncrossing trajectories will be completely discussed in Section IX.

Some examples of the results obtained by E. C. Stoner and D. R. Hartree are shown on the next figures. Figure 9 is taken from Stoner's report C.V.D. Mag. 16 (May, 1942) and gives the electron trajectories in a cylindrical magnetron when a constant voltage is suddenly applied on the anode. This means that the current is temperature limited at the beginning (*saturation current*). The anode radius r_a is twice the cathode radius r_c, and the notations are

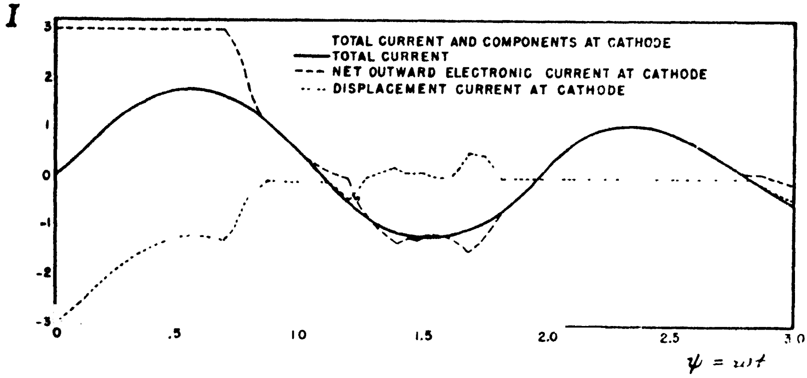


FIG. 10

$$S = \frac{r}{r_a} \quad \text{ordinate} \tag{76}$$

$$a = \frac{4\pi^2 eV_a}{mw^2 r_a^2} \qquad w = \frac{eM}{mc} = 2\omega_H \quad \text{EM units nonrational} \tag{77}$$

$$b = \frac{4\pi}{wV_a} \tag{78}$$

$$\delta = \frac{wt}{2\pi} \quad \text{abscissa} \tag{79}$$

The curves are labeled with the initial values δ_0 of δ. The figure corresponds to $a = 2$ and $bi_s = 3$ where i_s is the *saturation* current. Trajectories cross each other, and looking at the first few ones one may notice that the crossings with neighbor curves occur just before

$$\delta - \delta_0 \leq 1 \tag{80}$$

which means

$$2\omega_H(t - t_0) \leq 2\pi \tag{81}$$

Figure 10 shows the total current as a function of time. It should be noted that the saturation current is obtained up to $\delta = 0.7$, after which the current is space charge limited. During the time intervals when the current is negative $(1.1 < \delta < 2)$ some electrons fall back on the cathode and a displacement current is obtained on account of the positive field on the cathode

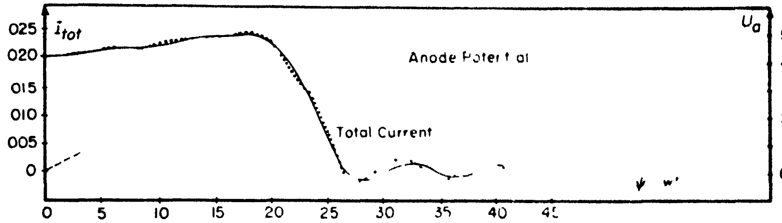


FIG. 11

Figure 11 is from Hartree's report C.V.D. Mag 23 (1942) and shows the anode voltage and the current as functions of ψ (which is the same as δ).

Figure 12 shows the trajectories. On account of the slow increase of the voltage, saturation is never reached. The ratio r_a/r_c is 2.718. Here $S = r/r_c$ instead of r/r_a.

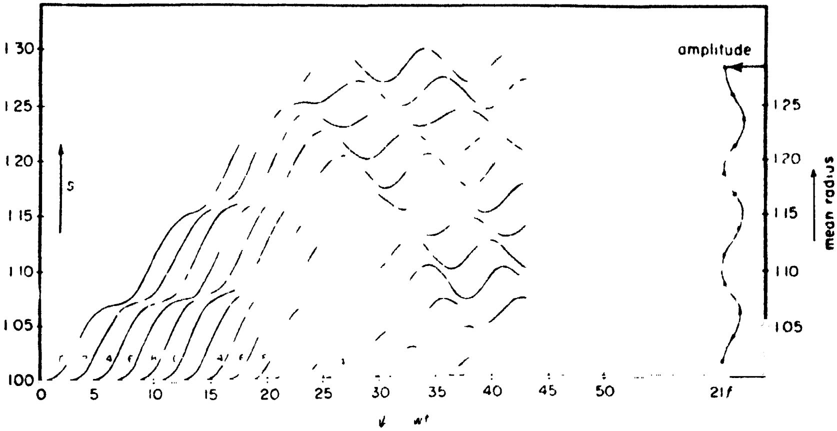


FIG. 12

Figures 13 and 14 are similar but for a still slower increase of voltage. In both cases most trajectories do not cross each other, and as in our former examples, some trajectories come very close and build up lines on which the space charge would be very large (perhaps infinite). However, just near the cathode, some intercrossings may develop, a point which will be discussed later The final voltage is constant and results in an oscillating current, which shows no sign of damping

As noticed in Section VIII, Hartree's theory is substantially the same as developed in Section VI of this report, but for the fact that the electric field on the cathode is not assumed to be always naught Let us call it $E_0(t)$ and rewrite eq 16

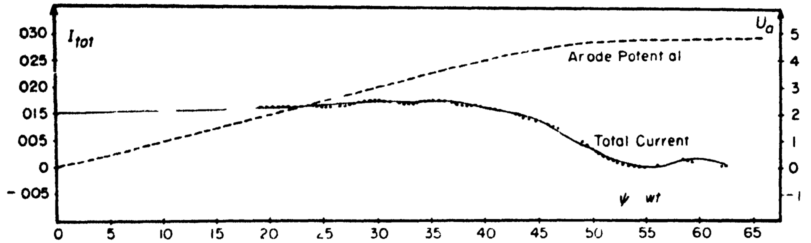


FIG. 13

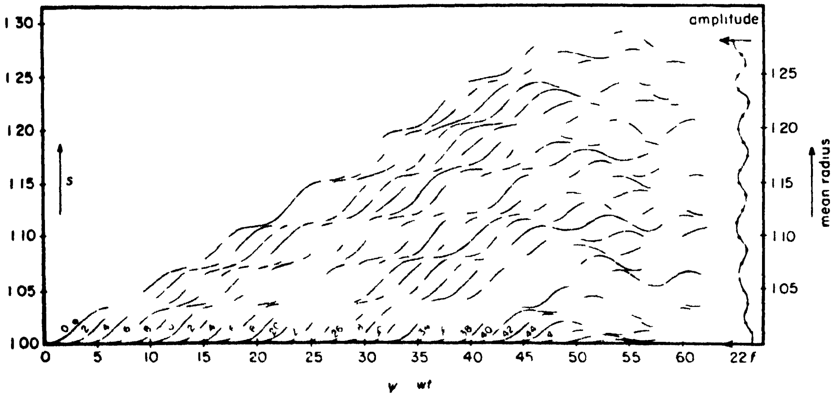


FIG. 14

$$E(t_0\, t) = V(t) - V(t_0) + E_0(t_0) \tag{82}$$

$$V(t) = -\frac{1}{\epsilon_0}\int_0^t I\,dt \tag{83}$$

$$I(t) = -\rho v - \epsilon_0 \frac{\partial E}{\partial t} \quad \text{total current (42)}$$

At the instant t_0 when an electron is emitted $E_0(t_0)$ is either zero (space charge limited current) or negative, yielding a positive force eE_0 which pulls all electrons out of the cathode (saturation current) There is no emission when E_0 is positive.

Let us choose our time origin $t = 0$ at the instant when the magnetron's operation is started. For $t = 0$ there is no charge in the tube, and the total charge introduced at time t is

$$q(t) = \epsilon_0 F(t) < 0 \tag{84}$$

A fraction $q - \epsilon_0 E_0(t)$ of this charge is actually found inside the space charge while $\epsilon_0 E_0(t)$ represents the charge induced by the field E_0 on the surface of the cathode. This induced charge is positive or negative according to the sign of E_0 (negative for saturation current positive during most of the intervals when electrons are falling back on the cathode). The case of space charge limited current (Section VI) was much simpler, since E_0 was always assumed zero.

The potential $V(d,t)$ of the anode is obtained from eq. 51, which Hartree integrates by parts:

$$\begin{aligned} V(d,t) &= -\int_{y=0}^{y=d} E(y,t)dy = -E(y,t)y\Big|_0^d + \int_{y=0}^{y=d} y\,dE \\ &= -E(d,t)d + \int_{y=0}^{y=d} y\left(\frac{\partial E}{\partial y}\right)_t dy \\ &= -E(d,t)d + \frac{1}{\epsilon_0}\int_0^d y\rho(y,t)dy \end{aligned} \tag{85}$$

by Poisson's relation. This integration by parts has the advantage of taking care of the field in vacuum, between the border of the space charge and the anode. *Assuming electrons never to reach the anode*, that field is uniform and equal to the field on the border of the space charge, namely (eq. 82),

$$E(d,t) = F(t) + E_0(0) \tag{86a}$$

since the electrons on the border left the cathode at $t = 0$. If we further assume *no saturation at the beginning*, $E_0(0) = 0$

$$E(d,t) = F(t) = \frac{1}{\epsilon_0} q(t) \tag{86b}$$

and we obtain

$$\epsilon_0 V(d,t) = -q(t)d + \int_0^d y\rho(y,t)dy \tag{87}$$

This is the translation, for the plane magnetron problem, of a formula used by Hartree.

A discussion of the physical meaning of this formula and of its application to multiple stream motions will be found in Section XVI.

IX. A GENERAL DISCUSSION OF ELECTRON TRAJECTORIES IN A PLANE MAGNETRON

The aim of this section is to discuss the conditions under which trajectories may cross each other or not, and to establish the limits of validity of the "single stream" motion, which was a basic assumption for the introduction of Llewellyn's formula (44) in Section VI.

The question was discussed for the case of *space charge limited* current, and it was proved that the trajectories of two electrons emitted at t_0 and $t_0 + dt_0$ could never cross each other. Let us now examine that same problem for *temperature limited current*, when there is a field $E_0(t_0)$ on the cathode and eq. 46 must be replaced by eq. 82. The field acting on the electron at time t is

$$E(t_0,t) = F(t) - F(t_0) + E_0(t_0) \tag{82}$$

Let us again define a function $f(t)$ by

$$\ddot{f} + 4\omega_H{}^2 f = \frac{e}{m} F \tag{48}$$

The position y of the electron is a solution of

$$y + 4\omega_H{}^2 y = \frac{e}{m} E \tag{41}$$

with the initial conditions

$$y = 0 \qquad \dot{y} = 0 \qquad \text{when } t = t_0 \tag{47}$$

Instead of eq. 49 the solution now reads

$$y(t_0,t) = f(t) - \frac{e}{m4\omega_H{}^2}[F(t_0) - E_0(t_0)] + C_1 \cos 2\omega_H(t - t_0) \quad . \\ + S_1 \sin 2\omega_H(t - t_0) \tag{88}$$

with

$$C_1 = C - \frac{e}{m4\omega_H{}^2} E_0 = -f(t_0) + \frac{e}{m4\omega_H{}^2}[F(t_0) - E_0(t_0)] \tag{89}$$

$$S_1 = S = -\frac{1}{2\omega_H}\dot{f}(t_0) \tag{90}$$

As in eq. 53 we want to compute

$$\left(\frac{\partial y}{\partial t_0}\right)_t = -\frac{e}{m4\omega_H{}^2}(\dot{F}_0 - \dot{E}_0) + \cos 2\omega_H(t - t_0)\left(\frac{\partial C_1}{\partial t_0} - 2\omega_H S_1\right) \\ + \sin 2\omega_H(t - t_0)\left(\frac{\partial S_1}{\partial t_0} + 2\omega_H C_1\right) \tag{91}$$

but

$$\frac{\partial S_1}{\partial t_0} + 2\omega_H C_1 = -\frac{e}{m2\omega_H} E_0 \tag{92}$$

$$\frac{\partial C_1}{\partial t_0} - 2\omega_H S_1 = \frac{e}{m4\omega_H^2} (\dot{F}_0 - \dot{E}_0) \tag{93}$$

Hence

$$\left(\frac{\partial y}{\partial t_0}\right)_t = \frac{e}{m4\omega_H^2} (\dot{F}_0 - \dot{E}_0)[-1 + \cos 2\omega_H(t - t_0)] - \frac{e}{m2\omega_H} E_0 \sin 2\omega_H(t - t_0) \tag{94}$$

a result which differs from eq. 56 by the E_0 terms Here we make use of eq. 42 and note

$$\dot{F}_0 - \dot{E}_0 = -\frac{1}{\epsilon_0}(I(t_0) + \epsilon_0 \dot{E}_0) = -\frac{i(t_0)}{\epsilon_0} \tag{95}$$

where $i(t_0)$ is the *electronic* part of the *current* leaving the cathode at time t_0.

If there is *no back current on the cathode at t_0* but only electrons leaving the cathode, $i_0(t_0)$ is positive and

$$\left(\frac{\partial y}{\partial t_0}\right)_t = \frac{e i_0}{\epsilon_0 m 4\omega_H^2} [1 - \cos 2\omega_H\tau] - \frac{e}{m2\omega_H} E_0 \sin 2\omega_H\tau$$
$$\text{transit time } \tau = t - t_0 \tag{96}$$

The field E_0 on the cathode is negative, and the force eE_0 acting on electrons is positive when saturation is obtained. Hence, for small τ, $(\partial y/\partial t_0)_t$ is negative, but it may become *positive* later on, when

$$1 - \cos 2\omega_H\tau < \frac{2\omega_H}{i_0} \epsilon_0 E_0 \sin 2\omega_H\tau \quad E_0 < 0 \tag{97}$$

a condition which is satisfied for $2\omega_H\tau \leq 2\pi$ when $1 - \cos 2\omega_H\tau$ is a small positive quantity and the sine is negative.

Trajectories obtained under conditions of *space charge limitation* ($E_0 = 0$) *do not cross each other*, but with *saturation conditions* trajectories *will cross just before $2\omega_H\tau$ equals 2π.*

This is exactly what is observed on Fig. 9 where trajectories cross each other when $\delta = 2\omega_H\tau/2\pi \leq 1$. As soon as this happens the single stream theory ceases to apply and the more complicated equations of Stoner must be used.

This checks very well with the results obtained in Sections VI and VII where no crossings were observed with space charge limited currents.

As for Hartree's problem (Section VIII), it needs a more detailed discussion.

All we have discussed is the question of trajectories crossing each other when electrons have been emitted at a very short interval of time *dt*. But it may happen that electronic emission stops for a finite time interval during which electrons fall back on the cathode (negative current). When electronic emission starts again afterward, it is perfectly feasible that the new trajectories may cross those of electrons emitted before the interval. As a matter of fact, this almost happens on Hartree's curves* (Fig. 14, at the bottom, curves 46, 48, 50 and Fig. 12, curves 24, 30). There is nothing to prevent electrons from leaving the cathode at a time when some of the previously emitted electrons are still falling back upon it!

X. STEADY PROBLEM: NEGATIVE RESISTANCE FOR VERY LOW FREQUENCIES

The problem of the plane magnetron operated under steady conditions was discussed in Section IV, Fig. 3. From formulas (24), (26), and (28) we obtain the relation between reduced current C and reduced voltage U

$$
\begin{aligned}
U &= \frac{2\epsilon_0}{M}\frac{V}{4\omega_H{}^2d^2}, \qquad M = -\frac{m\epsilon_0}{e} > 0\\
C &= \frac{2}{M}\frac{9J}{32\omega_H{}^3d} = \frac{4.5}{\psi_0 - \sin\psi_0}\\
U &= 1 + \left[\frac{C}{4.5}(1-\cos\psi_0)\right]^2 \qquad \psi_0 = 2\omega_H\tau,\ \tau,\ \text{transit time}
\end{aligned}
\tag{98}
$$

and we compute the static internal resistance of the magnetron

$$
\begin{aligned}
R &= \frac{dV}{dJ} = \frac{9d}{8\epsilon_0\omega_H}\frac{dU}{dC}\\
\frac{dU}{dC} &= \frac{2C}{(4.5)^2}(1-\cos\psi_0)\left[1-\cos\psi_0 + C\sin\psi_0\frac{d\psi_0}{dC}\right]
\end{aligned}
$$

From the C formula above we derive

$$dC(\psi_0 - \sin\psi_0) + C\,d\psi_0(1-\cos\psi_0) = 0$$

hence $d\psi_0/dC$, and

$$
\begin{aligned}
\frac{dU}{dC} &= \frac{2C}{(4.5)^2}[(1-\cos\psi_0)^2 - \sin\psi_0(\psi_0 - \sin\psi_0)]\\
&= \frac{2C}{(4.5)^2}[2 - 2\cos\psi_0 - \psi_0\sin\psi_0]
\end{aligned}
$$

---

* This could not happen on the examples discussed in Section VII because the current never went negative.

and finally

$$R = \frac{J}{M\epsilon_0(2\omega_H)^4}[2 - 2\cos\psi_0 - \psi_0\sin\psi_0] \tag{99}$$

This expression becomes negative along the branches P_1Q_2, P_2Q_3, $\cdots$ of Fig. 3 which correspond to a negative bracket in eq. 99.

$$P_1Q_2 \text{ branch } 2\pi \leq \psi_0 < \psi_2, \quad \psi_2 < 3\pi$$
$$P_2Q_3 \text{ branch } 4\pi \leq \psi_0 < \psi_3, \quad \psi_3 < 5\pi$$

and so on. Optimum conditions for negative resistance are obtained, for low frequencies, near

$$\psi_0 = 2n\pi + \frac{\pi}{2} \tag{99a}$$

in the middle of the P_nQ_{n+1} branches.

Looking at the P_1Q_2 branch on Fig. 3, we note a possibility for oscillations with

Average direct current	$\bar{C} \approx 0.635$
Average d-c voltage	$\bar{U} \approx 1.02$
Current oscillations	$2\Delta C = 0.72 - 0.55 = 0.17$
Voltage oscillations	$2\Delta U = 0.04$

hence an efficiency for low-frequency operation

$$\epsilon = \frac{1}{2}\frac{\Delta C \cdot \Delta U}{\bar{C} \cdot \bar{U}} = 0.0013 \tag{100}$$

that is very low.

XI. SMALL OSCILLATIONS OF HIGH FREQUENCY: FUNDAMENTAL EQUATIONS FOR THE TRAJECTORIES

The general method which will be followed was given in Section VI. It was shown that the single stream solution can be used, provided saturation is not reached, and the current never becomes negative. Both conditions will be assumed in this discussion. We start from the assumption of a current

$$I = I_0 + I_1 \cos\omega t = I_0(1 + a\cos\omega t), \quad a < 1, I_1 = aI_0 \tag{101}$$

with a constant I_0 term and superimposed oscillations. Imaginary exponentials will not be used in the computation since most of the formulas are nonlinear, which makes it safer to work only with real quantities. Referring to eq. 46, we introduce

$$F(t) = -\frac{1}{\epsilon_0}\int_0^t I dt = -\frac{I_0}{\epsilon_0}\left(t + \frac{a}{\omega}\sin\omega t\right) \tag{102}$$

and we write the equation for *self-consistent electron trajectories*, assuming saturation conditions never to be reached, hence the field E_0 on the cathode to remain zero. The integration proceeds exactly along the lines of eqs. 49 and 50, with

$$f(t) = \frac{I_0}{M4\omega_H^2}\left[t + \frac{a}{\omega}\,\frac{\sin\omega t}{1 - \left(\dfrac{\omega}{2\omega_H}\right)^2}\right] \tag{103}$$

and after computation of the C and S coefficients according to eq. 50, the net result is

$$\begin{aligned} y = \frac{I_0}{M8\omega_H^3}\Bigg[\psi_0 - \sin\psi_0 + a\Bigg\{&\frac{\frac{1}{\Omega}\sin\omega t}{1-\Omega^2} - \frac{1}{\Omega}\sin\omega t_0 \\ &- \frac{1}{1-\Omega^2}(\Omega\sin\omega t_0\cos\psi_0 + \cos\omega t_0\sin\psi_0)\Bigg\}\Bigg] \end{aligned} \tag{104}$$

with $\qquad \Omega = \omega/2\omega_H$

and $\qquad \psi_0 = 2\omega_H(t - t_0) \qquad (105)$

The first terms in $(\psi_0 - \sin\psi_0)$ represent the static solution (eq. 14), and the a terms yield the perturbation of the trajectory. It should be noted that the perturbation of y is purely linear. Nonlinear terms will be found only in the computation of the voltage. It is convenient to regroup terms in the a bracket:

$$\begin{aligned} \{\cdots\} = \frac{1}{\Omega(1-\Omega^2)}[&\sin\omega t - (1-\Omega^2)\sin\omega t_0 - \sin(\omega t_0 + \psi_0) \\ &+ (1-\Omega^2)\sin\omega t_0\cos\psi_0 + (1-\Omega)\cos\omega t_0\sin\psi_0] \end{aligned} \tag{106}$$

$$\{\cdots\} = \frac{1}{\Omega}\left[\frac{\sin\omega t - \sin(\omega t_0 + \psi_0)}{1-\Omega^2} + \sin\omega t_0(-1 + \cos\psi_0) + \frac{\cos\omega t_0\sin\psi_0}{1+\Omega}\right] \tag{107}$$

$$\{\cdots\} = \frac{1}{\Omega}\left[-\frac{2\sin\frac{\psi_0}{2}(1-\Omega)\cos\left(\omega\frac{t+t_0}{2} + \frac{\psi_0}{2}\right)}{1-\Omega^2} + \sin\omega t_0(-1+\cos\psi_0) + \frac{\cos\omega t_0\sin\psi_0}{1+\Omega}\right] \tag{108}$$

This expression shows that no term becomes infinite at *resonance* when

$$\omega = 2\omega_H \qquad \Omega = 1$$

The first term is 0/0 and a simple computation gives

$$\cdot\{\cdot\ \cdot\ \cdot\}_{\Omega=1} = \left[-\frac{\psi_0}{2}\cos 2\omega_H t + \sin 2\omega_H t_0(-1 + \cos \psi_0) \right.$$
$$\left. + \frac{1}{2}\cos 2\omega_H t_0 \sin \psi_0 \right] \quad (109)$$

$$y = \frac{I_0}{M8\omega_H{}^3}\,[\psi_0 - \sin\psi_0 + a\{\cdot\ \cdot\ \cdot\}] \quad (110)$$

These trajectories never cross each other, and the electronic motion retains the "single stream" type, provided $a < 1$. When $a > 1$ a special discussion is needed. In particular, the case of no direct current ($I_0 = 0$, $I_1 \neq 0$) presents all sorts of difficulties.

As stated before, nonlinear terms appear in the computation of the voltage (eq. 58).

$$V(y,t) = \frac{-1}{M4\omega_H{}^2}\int_{\theta=t_0}^{\theta=t} [F(t) - F(\theta)](1 - \cos\psi)I(\theta)d\theta \quad (111)$$

with $\psi = 2\omega_H(t - \theta)$.

Equation 111 yields the potential at point y, time t, provided t_0 and y are related together by eqs. 108 and 110. Using eqs. 101 and 102 we obtain

$$V = \frac{I_0{}^2}{\epsilon_0 M8\omega_H{}^3}\int_{\theta=t_0}^{\theta=t}\left[\psi + \frac{a}{\Omega}(\sin\omega t - \sin\omega\theta)\right]$$
$$(1 - \cos\psi)(1 + a\cos\omega\theta)d\theta \quad (112)$$

Assuming *small oscillations* we keep only terms linear in a.

$$V = \frac{I_0{}^2}{\epsilon_0 M8\omega_H{}^3}\int_{\theta=t_0}^{\theta=t}(1 - \cos\psi)$$
$$\left[\psi + a\left(\psi\cos\omega\theta + \frac{1}{\Omega}\sin\omega t - \frac{1}{\Omega}\sin\omega\theta\right)\right]d\theta \quad (113)$$

This, together with eqs. 108 and 110 builds the basis for our discussion.

XII. CHARACTERISTIC IMPEDANCE OF THE OSCILLATING PLANE MAGNETRON

The internal impedance of the oscillating magnetron is obtained by building the ratio

$$\frac{V_{osc}}{I_1} = \frac{V_{osc}}{aI_0} = \frac{1}{I_0}\frac{\partial V}{\partial a} = R\cos\omega t + X\sin\omega t \quad (114)$$

where V_{osc} represents the oscillating terms in V. There are terms in $\cos \omega t$ and $\sin \omega t$, and we may write down a complex impedance according to usual definitions by taking

$$Z = R + iX \tag{114a}$$

In computing $\partial V/\partial a$ we must remember that a enters the integral (113) in two different places: (a) in the lower limit t_0 according to eq. 110, (b) in the integral proper. According to our rule to keep only the terms linear in a, we shall drop the a terms in the integral when taking the derivative relative to the lower limit, hence

$$\frac{1}{I_0}\frac{\partial V}{\partial a} = \frac{I_0}{\epsilon_0 M 8\omega_H{}^3}\left[-(1 - \cos \psi_0)\psi_0\left(\frac{\partial t_0}{\partial a}\right) + \frac{1}{\Omega}\int_{t_0}^{t}(1 - \cos \psi)\right.$$
$$\left.[\Omega\psi \cos \omega\theta + \sin \omega t - \sin \omega\theta]d\theta\right]$$

The first term in the bracket represents the derivative with respect to the lower limit t_0, and we obtain $(\partial t_0/\partial a)_y$ by differentiating eq. 110, since y must be kept constant while a and t_0 vary simultaneously

$$-2\omega_H(1 - \cos \psi_0)\left(\frac{\partial t_0}{\partial a}\right)_y + \{\cdot\ \cdot\ \cdot\} = 0$$

since

$$\frac{\partial \psi_0}{\partial a} = -2\omega_H \frac{\partial t_0}{\partial a} \tag{115}$$

where the bracket $\{\cdot\ \cdot\ \cdot\}$ is given by eq. 108 and represents the coefficient of a in eq. 110. This yields

$$\frac{1}{I_0}\frac{\partial V}{\partial a} = \frac{I_0}{\epsilon_0 M 8\omega_H{}^3}\left[-\frac{\psi_0}{2\omega_H}\{\cdot\ \cdot\ \cdot\} + \frac{1}{\Omega}\int_{t_0}^{t}(1 - \cos \psi)\right.$$
$$\left.[\Omega\psi \cos \omega\theta + \sin \omega t - \sin \omega\theta]d\theta\right], \quad \Omega\psi = \omega(t - \theta) \tag{116}$$

The computation of the integral, if somewhat lengthy, does not offer any special difficulty and yields

$$\begin{aligned}\int_{t_0}^{t}(\cdot\ \cdot\ \cdot)[\cdot\ \cdot\ \cdot]d\theta = \Bigg|&(t - \theta)\sin \omega\theta + \theta \sin \omega t + \frac{1}{2\omega_H}\sin \omega t \sin \psi \\ &+ \frac{\omega_H}{(2\omega_H - \omega)^2}\cos(\psi + \omega\theta) + \frac{\omega(t - \theta)}{2(2\omega_H - \omega)} \\ &\sin(\psi + \omega\theta) - \frac{\omega_H}{(2\omega_H + \omega)^2}\cos(\psi - \omega\theta) \\ &+ \frac{\omega(t - \theta)}{2(2\omega_H + \omega)}\sin(\psi - \omega\theta)\Bigg|_{\theta = t_0}^{\theta = t}\end{aligned} \tag{117}$$

where $\psi = 2\omega_H(t - \theta)$

$\psi_0 = 2\omega_H(t - t_0) = 2\omega_H\tau$, τ transit time

This result is easily checked by direct differentiation of the right-hand terms. Substituting in eq. 116 one notices many term compensations between the two groups of terms.

Finally,

$$\frac{1}{I_0}\frac{\partial V}{\partial a} = R\cos\omega t + X\sin\omega t$$

$$= \left(\frac{I_0}{\epsilon_0 M 16\omega_H{}^4}\right)\left[-\left(\frac{\omega\tau}{1 - \Omega^2} + \frac{\sin\psi_0}{\Omega}\right)\sin\omega t + \frac{\cos\omega t}{2\Omega}\left[\frac{1}{(1 - \Omega)^2} - \frac{1}{(1 + \Omega)^2}\right] - \frac{\cos(\psi_0 + \omega t_0)}{2\Omega(1 - \Omega)^2} + \frac{\cos(\psi_0 - \omega t_0)}{2\Omega(1 + \Omega)^2}\right]$$

but

$$\psi_0 \pm \omega t_0 = 2\omega_H(t - t_0) \pm \omega t_0 = \pm\omega t + (2\omega_H + \omega)\tau = \pm\omega t + \psi_0(1 \mp \Omega),$$
$$\omega\tau = \Omega\psi_0$$

hence, after some elementary transformations

$$R = \left(\frac{I_0}{\epsilon_0 M 16\omega_H{}^4}\right)\frac{1}{2\Omega}\left[\frac{1 - \cos(1 - \Omega)\psi_0}{(1 - \Omega)^2} - \frac{1 - \cos(1 + \Omega)\psi_0}{(1 + \Omega)^2}\right]$$
$$X = \left(\frac{I_0}{\epsilon_0 M 16\omega_H{}^4}\right)\left[-\frac{\Omega\psi_0}{1 - \Omega^2} - \frac{\sin\psi_0}{\Omega} + \frac{\sin(1 - \Omega)\psi_0}{2\Omega(1 - \Omega)^2} + \frac{\sin(1 + \Omega)\psi_0}{2\Omega(1 + \Omega)^2}\right] \tag{118}$$

where $M = -\frac{m\epsilon_0}{e}$

$$\Omega = \frac{\omega}{2\omega_H}$$

$$\omega_H = -\frac{1}{2}\mu_0\frac{e}{m}H$$

$$\psi_0 = 2\omega_H(t - t_0)$$

These formulas completely describe the electrical properties of the one anode plane magnetron for small oscillations at all frequencies. It is obvious from the formulas that R is an even function of ω and X an odd function of ω.

XIII. MAGNETRON IMPEDANCE FOR LOW FREQUENCIES

Low frequencies are defined as

$$\omega \ll 2\omega_H, \quad \Omega \ll 1 \tag{119}$$

hence
$$\frac{1}{(1 \pm \Omega)^2} = 1 \mp 2\Omega \cdot \cdot \cdot$$

and eq. 118 yields

$$R = \left(\frac{I_0}{\epsilon_0 M 16\omega_H{}^4}\right)\left[\frac{\cos(1 + \Omega)\psi_0 - \cos(1 - \Omega)\psi_0}{2\Omega} + 2 - \cos(1 + \Omega)\psi_0 - \cos(1 - \Omega)\psi_0\right]$$

$$R = \left(\frac{I_0}{\epsilon_0 M 16\omega_H{}^4}\right)\left[-\frac{\sin\psi_0 \sin\omega\tau}{\Omega} + 2 - 2\cos\psi_0 \cos\omega\tau\right], \quad \psi_0 = 2\omega_H\tau$$

The transit time τ is finite while ω goes down to zero hence

$$\sin\omega\tau \approx \omega\tau$$
$$\cos\omega\tau \approx 1$$

and

$$R = \left(\frac{I_0}{\epsilon_0 M 16\omega_H{}^4}\right)[-\psi_0 \sin\psi_0 + 2 - 2\cos\psi_0] \qquad (120)$$

Since $\omega\tau/\Omega = \psi_0$. This is exactly the result previously obtained from the static characteristic (Section X, eq. 99). The reactance X can also be computed from eq. 118 for low frequencies. Keeping terms in Ω we find, after similar reductions

$$X = \left(\frac{I_0}{\epsilon_0 M 16\omega_H{}^4}\right)\omega\tau(-1 + \tfrac{1}{2}\psi_0 \sin\psi_0 + 2\cos\psi_0 \cdot \cdot \cdot) \qquad (121)$$

This is proportional to ω but depends in a rather complicated way upon the direct current I_0 through ψ_0. The bracket oscillates with increasing amplitude when ψ_0 increases, and takes positive and negative values. The zeros of X are just below $2n\pi$ and $(2n + 1)\pi$, as can be easily seen from the fact that $\frac{1}{2}\psi_0 \sin\psi_0$ represents the main term. The reactance X for low frequencies (eq. 121) takes rather large positive values for the values $\psi_0 = (2n + \frac{1}{2})\pi$ which yield negative resistance.

XIV. MAGNETRON IMPEDANCE FOR HIGH FREQUENCIES

High frequencies correspond to the neighborhood of Larmor frequency:

$$\omega \approx 2\omega_H \qquad \Omega = \frac{\omega}{2\omega_H} \approx 1 \qquad (122)$$

Looking at eq. 118 for the internal resistance of the magnetron, we note that the first term in $1 - \Omega$ is always positive whereas the second one in $1 + \Omega$ contributes to negative resistance.

$$R = \left(\frac{I_0}{\epsilon_0 M 16\omega \cdot \omega_H{}^3}\right)\left[\frac{1 - \cos\,(1 - \Omega)\psi_0}{(1 - \Omega)^2} - \frac{1 - \cos\,(1 + \Omega)\psi_0}{(1 + \Omega)^2}\right],$$

$$-e > 0 \text{ electrons, } I_0 > 0,\ \psi_0 = 2\omega_H\tau,\ \tau, \text{ transit time} \qquad (123)$$

The first term becomes very large when Ω approaches 1, hence if we want to obtain negative resistance, we must make it zero, while increasing the second term.

Negative resistance:

$$\begin{aligned} (1 - \Omega)\psi_0 &= 0,\ \pm 2\pi \cdots \pm 2n\pi \\ (1 + \Omega)\psi_0 &\approx \pi,\ \pm 3\pi \cdots \pm(2p + 1)\pi \end{aligned} \qquad (124)$$

The solution in $(1 - \Omega)\psi_0 = 0$ does not work, since the first term is 0/0 and not zero, and its actual value becomes $\frac{1}{2}\psi_0{}^2$, a finite quantity. The solutions of eq. 124 are

$$\psi_0 = (n + p + \tfrac{1}{2})\pi \qquad \Omega\psi_0 = (-n + p + \tfrac{1}{2})\pi \qquad (125)$$

with *positive and negative n values*, hence

$$\Omega = \frac{-n + p + \frac{1}{2}}{n + p + \frac{1}{2}}, \quad n, p \text{ integer} \qquad (126)$$

This yields all the possible Ω values giving negative resistance. High frequencies (eq. 122) are obtained for $|n| \ll p$, p large, for instance $n = \pm 1$, p large:

$$n = 1 \qquad \Omega = \frac{p - \frac{1}{2}}{p + \frac{3}{2}} < 1 \qquad (127a)$$

$$n = -1 \qquad \Omega = \frac{p + \frac{3}{2}}{p - \frac{1}{2}} > 1 \qquad (127b)$$

but the larger Ω the smaller the negative resistance, on account of the factor $1/(1 + \Omega)^2$ in eq. 123. For $\Omega \approx 1$ the negative resistance obtained under conditions (125) and (127) is

$$R \approx -\frac{1}{4}\frac{I_0}{\epsilon_0 M 16\omega_H{}^4} \qquad (128)$$

Assuming conditions (124) to (126) we compute the reactance

$$X = -\left(\frac{I_0}{\epsilon_0 M 16\omega_H{}^4}\right)\left[\frac{(-n + p + \frac{1}{2})(n + p + \frac{1}{2})^2}{2n(2p + 1)}\pi + (-1)^{n+p}\frac{n + p + \frac{1}{2}}{-n + p + \frac{1}{2}}\right] \qquad (129)$$

since $(1 - \Omega^2)\psi_0{}^2 = 2n\pi(2p + 1)\pi$.

For $\Omega \approx 1$ we must have $p \gg 1, |n|$, hence

$$X \approx -\left(\frac{I_0}{\epsilon_0 M 16\omega_H{}^4}\right)\left[\frac{p^2\pi}{4n}\right], \quad n \text{ positive or negative} \tag{130}$$

Negative resistance is thus connected with large reactance.

From the physical picture of the process, one might imagine that resonance conditions

$$\omega = 2\omega_H, \quad \Omega = 1 \tag{131}$$
$$\lambda H = 10655, \quad (\lambda \text{ centimeters and } H \text{ oersted})$$

should correspond to high efficiency for the one-anode plane magnetron, but this view is not confirmed by the theory.

Our theoretical investigations were limited to small oscillation $I_1 < I_0$ with a nonvanishing direct current I_0. They do not show any specially favorable conditions corresponding to resonance. The question of high efficiency is actually connected with large oscillations, where $a = I_1/I_0 > 1$, a condition which is contrary to the assumptions made about eq. 113 and which may even impair the whole method, since it does not seem to be compatible with the single stream assumption and the use of Llewellyn's formula (Section IV). The general discussion given in Section IX does not leave much hope for the validity of the single stream solution when the current becomes negative during certain time intervals.

As shown in eq. 125, negative resistance is obtained for

$$\psi_0 = k\pi + \frac{\pi}{2}, \quad k = n + p \quad \text{integer}$$

For low frequencies (see end of Section X, eq. 99*a*), half of these points drop out and only those corresponding to even k values remain.

XV. Discussion of Some Special Examples

A number of special cases were computed, in order to reach a better understanding of the behavior of the tube. Computations can be found in the original report,[10] and we shall only discuss the practical results, as shown on the drawings.

Figure 15 refers to a case of oscillations without direct current

$$I = I_1 \cos \omega t \tag{132}$$

starting at T_0 where $\omega T_0 = -\pi/2$. The lower curves show the variation of I and of its integral, the charge Q. The upper curves describe the motion of electrons, with y (distance from cathode) as a function of time. It was assumed that a certain amount of space charge (extending up to

$y = 2$) preexisted in the tube. These electrons give the trajectories above the B line. Trajectories below the B line correspond to electrons emitted from the cathode during the interval $-\pi/2 < \omega t < \pi/2$ and falling back upon it when $\pi/2 < \omega\tau < 3\pi/2$. Trajectories do not intercross each other during the first period, but a layer of intercrossing trajectories (shaded region) develops after the first time interval of negative current and should progressively expand afterward. This figure illustrates the fact pointed out at the end of Section IX, that negative currents lead to intercrossing trajectories. The figure corresponds to the case

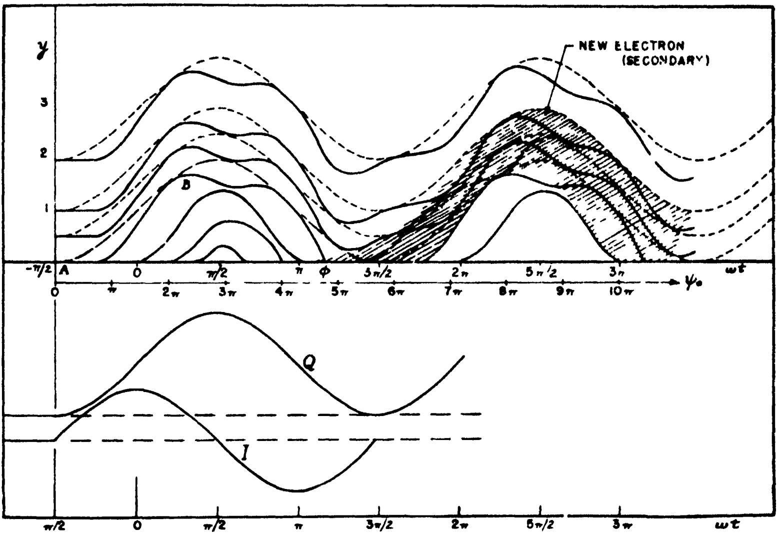


FIG. 15

$$\omega < 2\omega_H \qquad \omega \approx \cdot 7\omega_H \tag{133}$$

Figure 16 refers to a similar problem when exact resonance oscillations are used

$$\omega = 2\omega_H \qquad \Omega = 1$$

Electrons which were previously in the space charge between cathode and anode take larger and larger oscillations until they finally hit either cathode or anode. In the final stage one finds only electrons emitted and reabsorbed by the cathode. This means a positive resistance due to energy losses by cathode back bombardment.

Electron trajectories in this final stage will now be discussed. Three cases must be distinguished:

C Initially no space charge, and no charge on the anode at the start. Moderate amplitude of the oscillations I_1, electrons never being able to reach the anode

$$y_{\max} \leq d$$

D Initially no space charge, but anode initially charged negatively. Or also: anode initially uncharged but large I_1 oscillations, which bring some electrons to the anode during the first oscillation and charge it negatively for the rest of the operation.

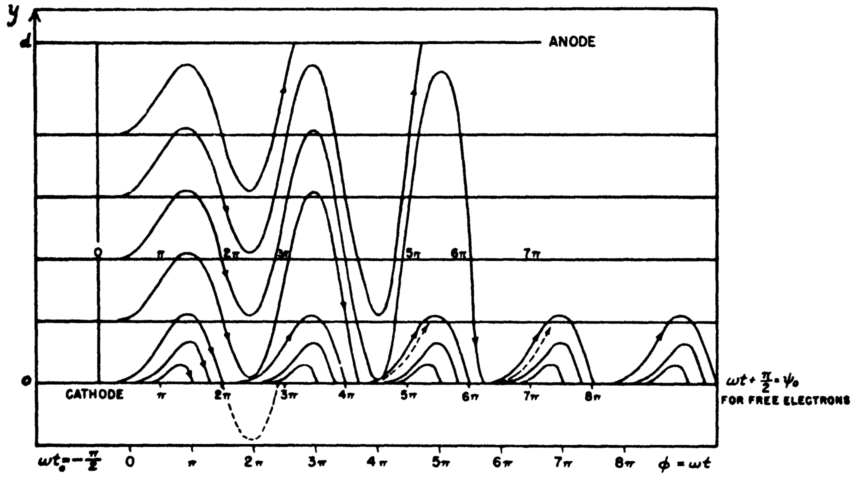


Fig. 16

In both cases C and D the motion is "single stream."

E Initial space charge, as shown on Fig. 16. Hence anode initially charged positively. Some electrons hit back the cathode, hence *double stream* motion, with positive charge on the anode.

Case C corresponds to *moderately large oscillations*, when electrons never reach the anode (anode in position C on Fig. 17). Electrons are emitted during the whole interval of positive current $(-\pi/2 < \phi < \pi/2)$ and reabsorbed during the negative current interval $(\pi/2 < \phi < 3\pi/2)$.

For a *large amplitude* I_1 one would find $y_{\max} > d$ (anode in position D on Fig. 17). During the first oscillations, some electrons reach the anode and give it a negative charge. This negatively charged anode creates a field retarding electron emission from the cathode. Instead of starting at $\phi = -\pi/2$, emission is delayed until $\phi = \Phi_0$ as shown on Fig. 17 (case D). These electrons strike back at Φ_1, and the cathode does not emit between Φ_1 and $2\pi + \Phi_0$, since during this time interval the negative charge on the anode induces a positive charge on the cathode, and the structure simply behaves like a condenser.

Similar oscillations and operation can be obtained even for small I_1 amplitude, if the anode has been initially charged negatively.

A remark should be made here to explain a mistake which must be avoided: one might be tempted, in the case of large I_1 oscillations, to assume an electronic emission starting at $\phi = -\pi/2$ on every oscillation thus giving each time a bunch of electrons reaching the anode. But it would be a serious mistake to imagine this electronic motion to represent a solution for oscillations I_1 superimposed to a small direct current (corresponding to the electrons reaching the anode). We have assumed from the beginning that there was no direct current $I_0 = 0$ and trajectories computed for Figs. 16 and 17 are valid only under this assumption.

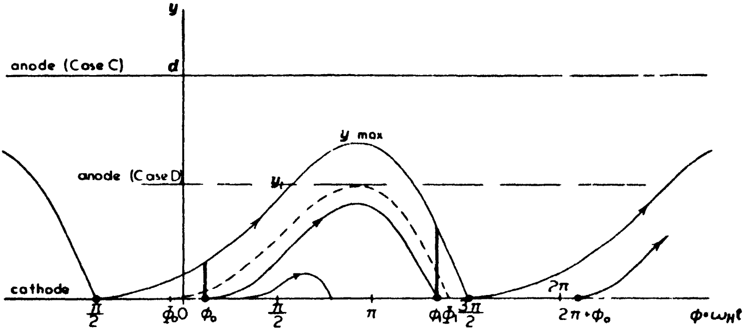


FIG. 17

A computation of the voltage proves that the *resistance* is always *positive*. The average anode voltage V_a is below cutoff, and depends upon the amplitude of oscillations I_1 (Fig. 18). Case C corresponds to a curve starting from $V = 0$, $I_1 = 0$ and rising up to a point I_{1m} when electrons would reach the anode. A typical D curve starts from $V = 0$ at a certain I_{10} amplitude of oscillations and rises up to the D limit curve when electrons reach the anode.

In an oscillating magnetron (not self-sustained in our case!) electrons may reach the anode for voltages much below cutoff.

A possibility for self-sustained oscillations can be found only when a direct current I_0 is added to the I_1 oscillations, since such a condition is necessary to draw some energy from the d-c supply battery connected to the anode.

Case E involves double stream motions and is discussed in the next section.

XVI. DOUBLE STREAM ELECTRONIC MOTIONS: GENERAL FORMULAS

Up to now, double stream motion was considered only in the steady case (Section V) when it could easily be reduced to the single stream problem.

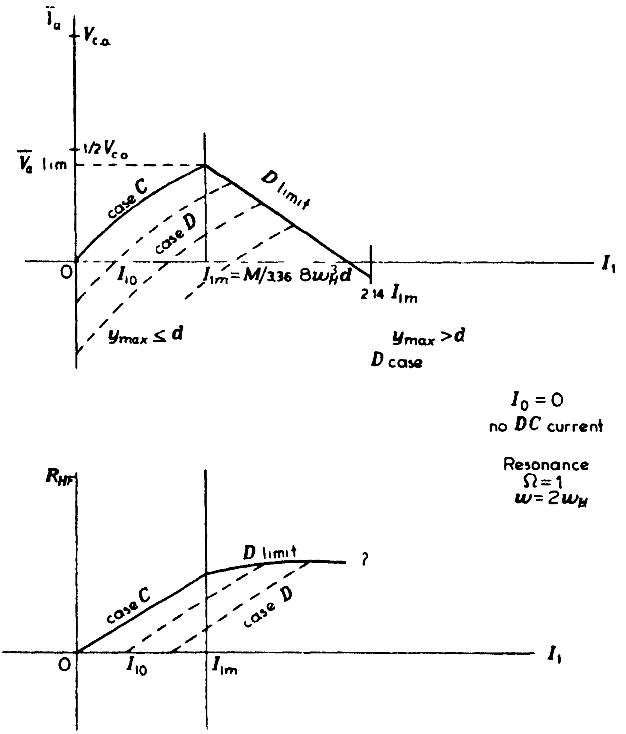


FIG. 18

Let us now consider a magnetron operated with variable total current $I(t)$. At a distance y from the cathode, the space charge density $\rho(y,t)$ can be divided into two parts:

$\rho_1(y,t)$ moving with velocity v_1

$\rho_2(y,t)$ moving with velocity v_2

$$\begin{aligned} \epsilon_0 \frac{\partial E}{\partial y} &= \rho = \rho_1 + \rho_2 \\ -I &= \rho_1 v_1 + \rho_2 v_2 + \epsilon_0 \frac{\partial E}{\partial t} \end{aligned} \tag{134}$$

Here again we assume a *plane space charge* distribution in the one-anode magnetron structure, and we state that $\rho_1\rho_2$ depend only upon y and t but not upon x.

According to Llewellyn's method (Section VI) we build the expression $(dE/dt)_1$ following the motion of the $\rho_1 v_1$ stream of electrons.

$$\left(\frac{dE}{dt}\right)_1 = \frac{\partial E}{\partial y}\left(\frac{dy}{dt}\right)_1 + \frac{\partial E}{\partial t} = \frac{1}{\epsilon_0}\left(\rho v_1 + \epsilon_0 \frac{\partial E}{\partial t}\right) \tag{135}$$

Hence

$$\left(\frac{dE}{dt}\right)_1 = -\frac{I}{\epsilon_0} + -\frac{\rho_2}{\epsilon_0}(v_1 - v_2) \tag{136}$$

following v_1 electrons, and in a similar way

$$\left(\frac{dE}{dt}\right)_2 = -\frac{I}{\epsilon_0} + \frac{\rho_1}{\epsilon_0}(v_2 - v_1) \tag{137}$$

following v_2 electrons.

The single stream solution of Llewellyn was

$$\frac{dE}{dt} = -\frac{I}{\epsilon_0}, \quad E = \frac{1}{\epsilon_0}\int_{t_0}^{t}(-I)dt$$

Integration of formula (136) yields:

$$E_1 = E_{10} + \frac{1}{\epsilon_0}\int_{t_0}^{t}(-I)dt + \frac{1}{\epsilon_0}\int_{t_0}^{t}\rho_2(v_1 - v_2)dt \tag{138}$$

where E_{10} = the field on the cathode at t_0 when v_1 electrons were emitted

$Q(t_0,t) = \int_{t_0}^{t}(-I)dt$ is the charge emitted from the cathode between t_0 and t

$$Q_2 = \int_{t_0}^{t}\rho_2(v_1 - v_2)dt$$

is the charge brought between our electrons and the cathode by the v_2 stream, since $(v_1 - v_2)$ represents the relative velocity of the v_1, v_2 streams. Let us call $Q(y,t)$ the total charge (per unit area) of the space charge between the electrons we are considering and the cathode. Assuming no saturation ($E_{10} = 0$), eq. 138 yields

$$E_1 = \frac{Q}{\epsilon_0} = \frac{Q(t_0,t) + Q_2}{\epsilon_0} = \frac{1}{\epsilon_0}\int_{t_0}^{t}(-I)dt + \frac{1}{\epsilon_0}\int_{t_0}^{t}\rho_2(v_1 - v_2)dt \tag{139}$$

This result has a very clear physical meaning and can be used also for a direct interpretation of Llewellyn's formula for single stream motion.

The situation is explained on Fig. 19, assuming a stream v_2 of electrons to strike back on the cathode during the emission of another stream v_1. Both streams cross at AB, $A'B'$, the charge Q below the v_1 trajectory is thus increased and so is the force eE_1 acting on the v_1 electrons. This results in a deflection, but most important of all is the fact that the additional force eE_1 *persists all along the trajectory.* Hence the main change consists in an additional constant acceleration. This is easily proved: the v_2 stream strikes the cathode between t_2t_2' and is *immediately reemitted,* since the total current $I(t)$ is our primary data, and must follow its regular course. Hence there is an additional Q_2 emission from t_2 to t_2', with trajectories spreading fanwise between the "deflected" and the "normal" single stream trajectories. The point is that the additional charge Q_2 cannot be lost, hence the force on the v_1 electrons is permanently increased by an $[e(Q_2/\epsilon_0)]$ term.

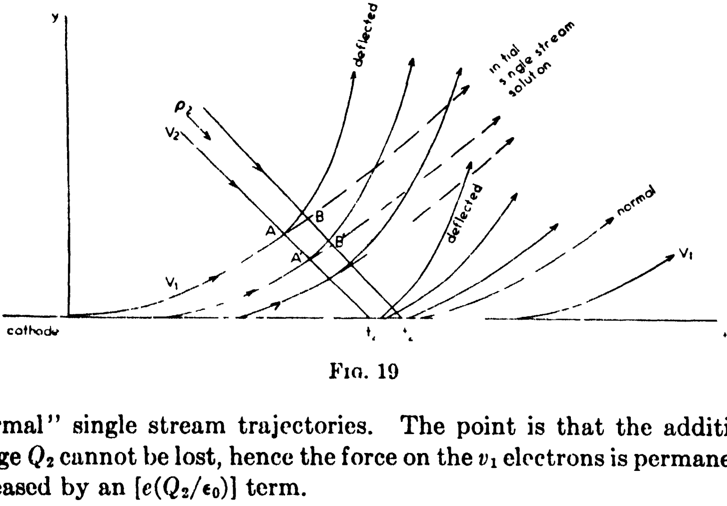


FIG. 19

With any type of trajectory, the anode voltage can always be computed from eqs. 51 and 85

$$V_a = -\int_0^d E(y,t)dy = -E_a d + \frac{1}{\epsilon_0}\int_{y=0}^{y=d} y\,dq \qquad (140)$$

where E_a = the field on the anode
$y = d$ and
$dq = \rho dy$ represents the space charge.

We may have a field E_0 on the cathode, and

$$E_a - E_0 = \frac{1}{\epsilon_0}\int dq \qquad (141)$$

E_0 is zero during the intervals of electronic emission (space charge limitation) but it may become positive ($eE_0 < 0$) when electrons strike back upon the cathode and the current is negative.

From eqs. 140 and 141 we obtain

$$V_a = -E_0 d - \frac{1}{\epsilon_0}\int_{y=0}^{y=d} (d-y)dq = -E_0 d - \frac{1}{\epsilon_0}\int_{\text{cathode}}^{\text{anode}} z dq \quad (142)$$

$$= -E_0 d + \frac{1}{\epsilon_0}\int_{\text{cathode}}^{\text{anode}} z|dq|$$

where $z = d - y$ is the *distance from the anode* of the layer with charge dq per unit area. The final result is obvious and could have been directly obtained by considering each negative layer dq and the equal positive charge $|dq|$ on the anode as a plane condenser of capacity ϵ_0/z per unit area, then adding each contribution to the voltage. Explained in simple words it means that charges dq are more effective in building up the anode voltage V_a when they are located near the cathode ($z \approx d$) rather than near the anode ($z \approx 0$).

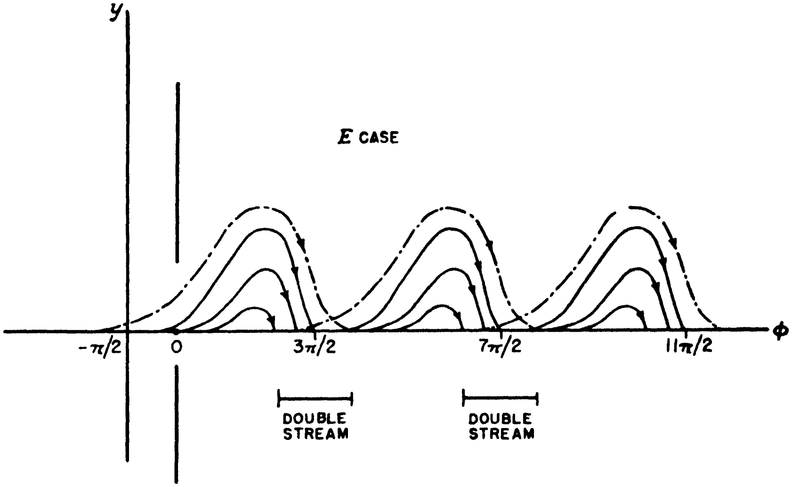


FIG. 20

We may now discuss the last case E of oscillations without direct current, which was stated in Section XV but could not be discussed with the single stream method. This is a problem of positively charged anode, with small oscillations and electrons never reaching the anode. The solution is obtained by adding an additional layer of electrons on top of the trajectories of case C, as shown on Fig. 20, where the additional trajectories have been drawn as dot and dash lines. The motion is "single stream" for the main part of the trajectories, with intercrossing trajectories and double stream just near the cathode around $\phi = -\pi/2$, $3\pi/2$, $7\pi/2$ · · · In these regions, the disposition of the trajectories is similar to that shown on Fig. 19. Such should be the motion for a small amount of added space charge, but it would become more and more complicated for large space charges and small oscillations.

The average voltage $\bar{V}_a$ is obviously higher than for the C case, and we may complete our schematic drawing of Fig. 18 as shown on Fig. 21 which is self-explanatory. The limiting curve of Fig. 18 can now be prolonged to the left until it reaches the $\bar{V}_a$ axis at the cutoff voltage V_{co},

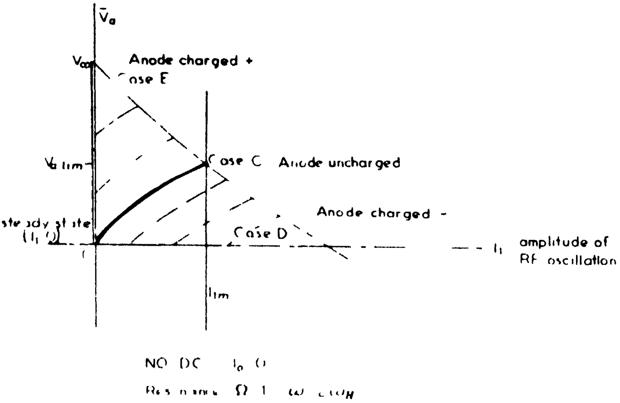


FIG 21

but it is hard to foresee whether the steady case obtained at the limit $I_1 \rightarrow 0$ will be the single stream B or the double stream A solution.

We have thus obtained a qualitative description of the different C, D, E, types of resonance oscillations in a magnetron operated without direct current. All these different types must be sustained from outside by a radio-frequency power supply, but the power needed is small and oscillations reach high amplitudes.

XVII. LARGE RESONANT OSCILLATIONS WITH MODERATE DIRECT CURRENT

The preceding discussion showed that no negative resistance could be obtained at exact resonance ($\omega = 2\omega_H$, $\Omega = 1$) for small oscillations. We wish now to discuss the case of rather large oscillations $I_0 = I_1$ ($a = 1$) when the single stream solution of Section XI can still be used.

Trajectories are given by eqs. 109 and 110 in the case of exact resonance.

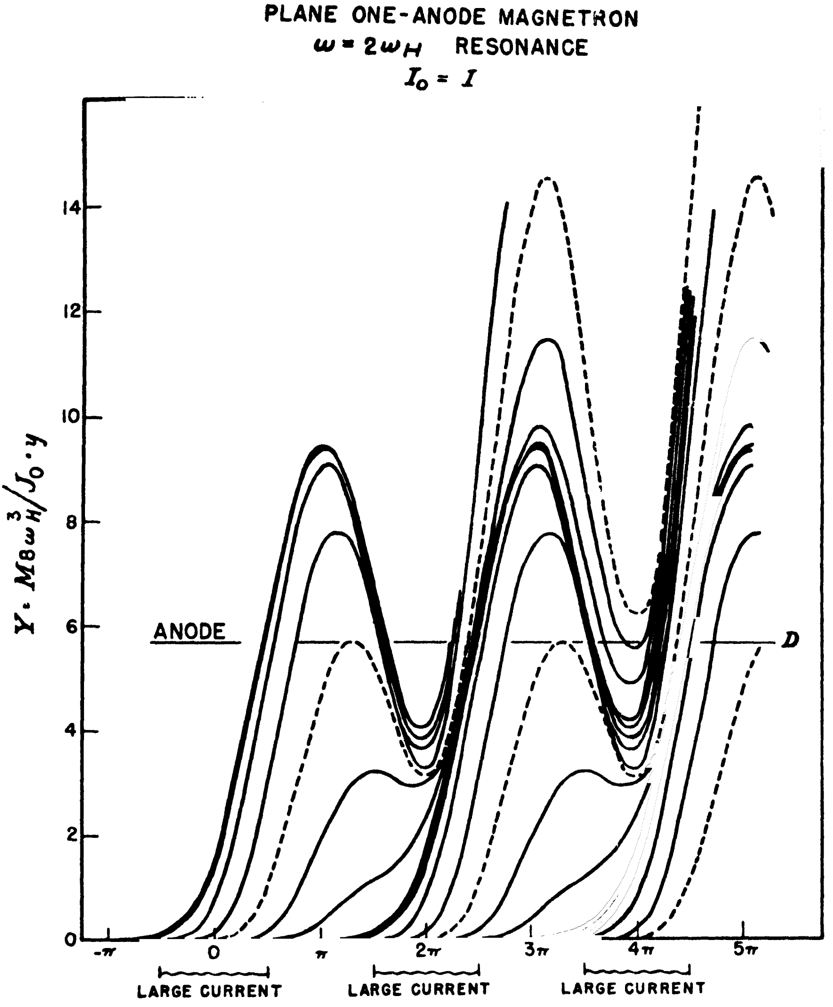


Fig. 22

The limiting case $a = 1$, $\Omega = 1$ is very interesting to consider and corresponding trajectories have been plotted on Fig. 22. They exhibit large oscillations of increasing amplitude and a very remarkable bunching effect, which increases as time goes on. Large electron emission is obtained during the intervals

$$-\frac{\pi}{2} < \phi_0 < \frac{\pi}{2} \qquad \frac{3\pi}{2} < \phi_0 < \frac{5\pi}{2} \cdots \qquad \phi_0 = \omega t_0 = 2\omega_H t_0 \tag{143}$$

and the trajectories in these bundles corresponding to $\phi_0 = 0, 2\pi, 4\pi \cdots$ have been drawn in dashed line.

Most favorable conditions for sustained oscillations should be found when anode bombardment is low. Hence we should attempt to locate the anode at a distance where the main trajectory has a maximum and its electrons reach the anode with zero $\dot{y}$ velocity. This is obtained at $Y = 5.7$. The main trajectory reaches this distance at $\phi = \pi + (\pi/4)$ with zero velocity, but the following electrons take a large swing and strike back on the anode with full speed between $2\pi + (3\pi/8)$ and $3\pi + (\pi/4)$. The actual result is not satisfactory. A computation of the voltage yields a positive radio-frequency resistance.

When time goes on, the electron trajectories exhibit larger and larger oscillations, and smaller terms (like sine or cosine terms) can be neglected in front of other terms increasing linearly with time. Keeping only these dominant terms in eqs. 109 and 110 we obtain

$$\begin{aligned} Y &= M\frac{8\omega_H{}^3}{J_0}y \\ &= (\phi - \phi_0)(1 - \tfrac{1}{2}\cos\phi) \cdots \\ \Omega &= 1, \qquad a = 1, \qquad \phi - \phi_0 \gg 1 \end{aligned} \tag{144}$$

Neglected sine and cosine terms contribute less than ±2.5 in the final result. An electron beam corresponding to a whole period is obtained by taking

$$-\frac{\pi}{2} < \phi_0 < \frac{3\pi}{2}$$

which gives an overall variation

$$\Delta Y = 2\pi(1 - \tfrac{1}{2}\cos\phi)$$

The average trajectory ($\phi_0 = 0$) oscillates between $Y = \frac{3}{2}\phi$ and $\frac{1}{2}\phi$. Taking ΔY as a measure of the amount of bunching we find

$$\left.\begin{aligned} Y &= \tfrac{3}{2}\phi \qquad \phi = (2k+1)\pi, \quad \Delta Y = 3\pi \quad \text{min. bunching} \\ Y &= \tfrac{1}{2}\phi \qquad \phi = 2k\pi, \quad \Delta Y = \pi \quad \text{max. bunching} \end{aligned}\right. \tag{145}$$

the curves of Fig. 23 visualize these constant results. Y increases with ϕ while the ΔY's keep constant values, which indicate an increase of the relative bunching, since $\Delta Y/Y$ decreases steadily.

We may again try to see what happens if the anode is located at one of the maxima of Y

$$\text{anode } D = \frac{M8\omega_H{}^3}{I_0} \qquad d = \left(3k + \frac{3}{2}\right)\pi, \quad \begin{cases} \phi_0 = 0 \\ \phi = (2k+1)\pi \end{cases} \tag{146}$$

but here again the magnetron exhibits a positive resistance. The large oscillations of electrons in front of the anode result in a succession of cavitations alternating with heavy bombardment on the anode. This means a large energy dissipation on the anode and explains the positive resistance obtained.

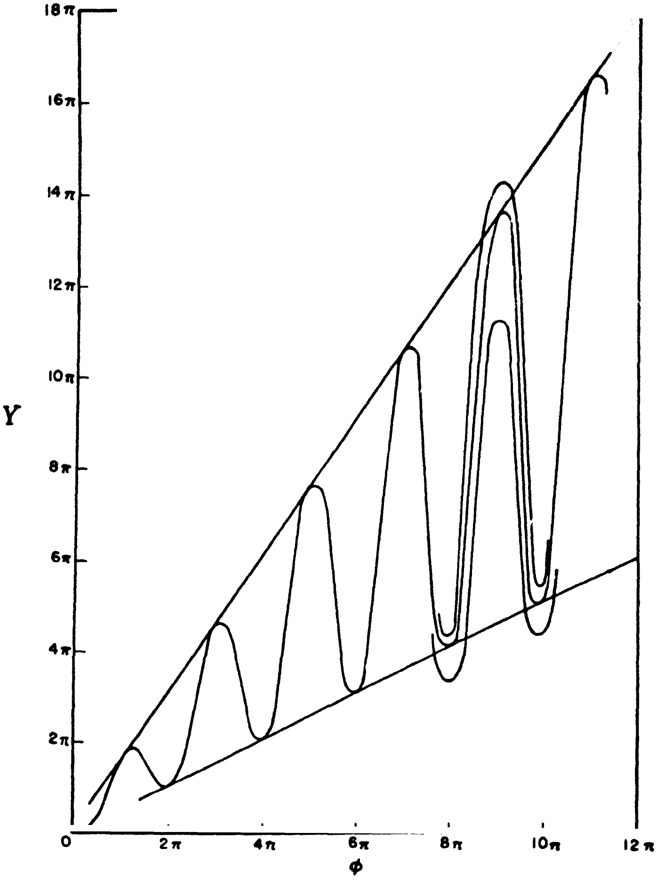


FIG. 23

XVIII. EFFICIENCY AND NEGATIVE RESISTANCE IN ONE-ANODE MAGNETRONS

In the discussion of efficiency for vacuum tubes, there is a well-known result.

Best efficiency is obtained when a virtual cathode is built just in front of the anode. The explanation is straight-forward. On a virtual cathode, electrons are slowed down, and their velocity drops to zero. Hence energy losses by anode bombardment are prevented.

It is very remarkable that we did not reach a similar conclusion in the discussion of the operation of magnetrons. As a matter of fact, the conditions for obtaining negative resistance correspond to large velocities for electrons impinging on the anode. For *low frequencies* (Sections X and XIII) negative resistance was found when

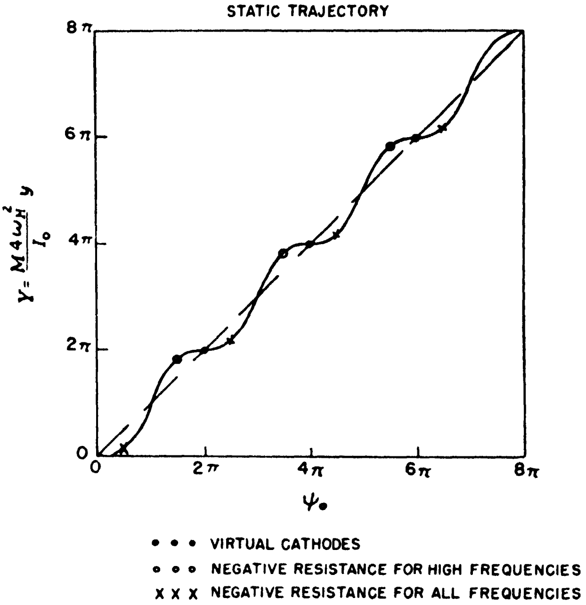


FIG. 24

$$\psi_0 \approx 2n\pi + \frac{\pi}{2} \tag{147}$$

while for *high frequencies* (Section XIV) the condition was

$$\begin{aligned} \psi_0 &= k\pi + \frac{\pi}{2} \\ \psi_0 &= 2\omega_H(t - t_0), \quad t - t_0, \text{ transit time} \end{aligned} \tag{148}$$

This does not correspond to virtual cathodes ($\psi_0 = 2n\pi$) on the static trajectory (see Fig. 24) but to a finite average electron velocity, since the steady average motion is

$$\begin{aligned} Y &= \frac{M8\omega_H^3}{I_0} y = \psi_0 - \sin \psi_0 \\ Y' &= \frac{1}{2\omega_H} \dot{Y} = \frac{M4\omega_H^2}{I_0} \dot{y} = 1 - \cos \psi_0 \end{aligned} \tag{149}$$

and both preceding conditions make $\cos \psi_0 = 0$, $\dot{y}$ finite.

This situation justifies a discussion of the physical meaning of the conditions resulting in negative resistance.

When the magnetron is operated under steady conditions (no oscillations), a current I_0 corresponds to a voltage V_0 and energy considerations prove that

$$-eV_0 = \tfrac{1}{2}mv^2 = \tfrac{1}{2}m(\dot{x}^2 + \dot{y}^2) = \tfrac{1}{2}m(4\omega_H^2 y^2 + \dot{y}^2) \tag{150}$$

where y is the anode distance and $\dot{x} = 2\omega_H y$ as shown in Section III. The number of electrons hitting the anode per second (and per unit area) is $(-1/e)I_0$; hence the energy dissipated per second by anode bombardment amounts to

$$-\frac{1}{e} I_0 \frac{1}{2} m(4\omega_H^2 y^2 + \dot{y}^2) = I_0 V_0 \tag{151}$$

The whole power supplied by the battery is used in heating the anode.

Now what happens when current oscillations

$$I_1 \cos \phi = I_0 a \cos \phi, \quad \phi = \omega t \tag{152}$$

are superimposed on the average direct current I_0? The voltage is represented by an expansion

$$\begin{aligned} V = V_0 + a(V_{1c} \cos \phi + V_{1s} \sin \phi) + a^2[V_{2c} \cos^2 \phi \\ + V_{cs} \cos \phi \sin \phi + V_{2s} \sin^2 \phi] \end{aligned} \tag{153}$$

and comparing this expression with eq. 114 we note

$$V_{1c} = RI_0 \qquad V_{1s} = XI_0 \tag{154}$$

Averaging over one period, we obtain

$$\bar{V} = V_0 + \frac{a^2}{2}(V_{2c} + V_{2s}) \tag{155}$$

which yields the d-c *power supply* from the battery

$$VI_0 = V_0 I_0 + \frac{a^2}{2} I_0 (V_{2c} + V_{2s}) = P_{\text{dc}} \tag{156}$$

On the other hand we obtain a similar expansion for the velocity of electrons hitting the anode,

$$\dot{y} = \dot{y}_0 + a(u_{1c} \cos \phi + u_{1s} \sin \phi) + a^2[u_{2c} \cos^2 \phi + u_{cs} \cos \phi \sin \phi + u_{2s} \sin^2 \phi] \cdots \tag{157}$$

hence

$$\dot{y}^2 = \dot{y}_0^2 + 2a\dot{y}_0(u_{1c} \cos \phi + u_{1s} \sin \phi) + a^2 \{(\cdots)^2 + 2\dot{y}_0[\cdots]\} \tag{158}$$

The amount of energy lost by anode bombardment is obtained by multiplying the kinetic energy $\frac{1}{2}mv^2$ by the number dn of electrons hitting the anode

$$P_{\text{bomb}} = \frac{1}{2} m \frac{1}{T} \int_T v^2 \, dn \text{ per second} \tag{159}$$

$$T = \frac{2\pi}{\omega} = \text{period}$$

For high frequencies, dn is a function of ϕ which must be computed from the trajectories but for *very low frequencies* (when the field on the anode is almost constant and displacement current on the anode can be neglected) we may write

$$dn = \frac{1}{|e|} I_0(1 + a \cos \phi)dt$$

and

$$P_{\text{bomb}} = \frac{mI_0}{2|e|} (4\omega_H{}^2y^2 + \dot{y}_0{}^2 + a^2 \overline{\{\cdots\}}) + \frac{m}{|e|} I_0\dot{y}_0u_{1c}a^2 \cos^2 \psi$$

$$= V_0I_0 + a^2 \overline{\{\cdots\}} \frac{mI_0}{2|e|} + \frac{m}{2|e|} I_0y_0u_{1c}a^2 \tag{160}$$

Negative resistance appears when the energy P_{bomb} dissipated by *anode bombardment* is *less than* the *d-c energy supplied* by the battery:

$$P_\omega = P_{\text{dc}} - P_{\text{bomb}} \tag{161}$$

the power P_ω thus available on the frequency ω appears through the mechanism of a negative resistance $R < 0$

$$P_\omega = \tfrac{1}{2}|R|I_1{}^2 = \tfrac{1}{2}|R|I_0{}^2a^2 \tag{162}$$

This is a rather indirect mechanism, which bears no precise connection with the formation of virtual cathodes. Hence *one-anode magnetrons should exhibit rather low efficiency* since negative resistance and virtual cathode conditions cannot be realized together.

These energetic considerations are clear enough and can be checked in the case of *low frequencies*, where computations are simpler on account of relation (160). One can prove that condition (147), which corresponds to negative resistance, also yields a positive value for P_ω in (151) and that both formulas for R and P_ω are identical.

For high frequencies, these considerations lead to condition (148).

XIX. PHYSICAL MEANING OF CONDITIONS FOR NEGATIVE RESISTANCE

The discussion of the preceding section showed the meaning of one of the conditions for negative resistance and emphasized the rather indirect process involved. The other condition will be found to have a very clear physical interpretation and to explain our failure to obtain negative resistance in the examples involving exact resonance (Sections XV and XVI). Let us first rewrite the conditions for negative resistance, as obtained in Section XIV

$$(1 - \Omega)\psi_0 = 2n\pi \qquad (1 + \Omega)\psi_0 = (2p + 1)\pi \tag{124}$$

$$\psi_0 = 2\omega_H(t - t_0) = (n + p + \tfrac{1}{2}\pi) \tag{125}$$

$$\Omega = \frac{\omega}{2\omega_H} = \frac{-n + p + \frac{1}{2}}{n + p + \frac{1}{2}} \tag{126}$$

We want to discuss the behavior of the trajectory, which is given by eqs. 108 and 110:

$$Y = \psi_0 - \sin\psi_0 + \frac{a}{\Omega}\left[-\frac{2}{1 - \Omega^2}\sin\frac{\psi_0}{2}(1 - \Omega)\cos\Phi \right.$$
$$\left. + \sin\phi_0(-1 + \cos\psi_0) + \frac{1}{1 + \Omega}\cos\phi_0\sin\psi_0\right] \tag{163}$$

with $\phi = \omega t$.

$$\Phi = \tfrac{1}{2}(\phi + \phi_0 + \psi_0) = \tfrac{1}{2}(\omega + 2\omega_H)t + \tfrac{1}{2}(\omega - 2\omega_H)t_0$$
$$\Phi = \frac{\phi}{2}\left(1 + \frac{1}{\Omega}\right) + \frac{\phi_0}{2}\left(1 - \frac{1}{\Omega}\right) \tag{164}$$

To visualize the results we choose a specific example

$$n = 1 \qquad p = 10 \qquad \Omega = \frac{9.5}{11.5} = 0.826$$
$$a = 1 \qquad \frac{a}{\Omega} = 1.21 \qquad \frac{2}{1 - \Omega^2} = 6.3 \tag{165}$$

These conditions correspond to negative resistance when $\psi_0 = 11.5\pi$.

We choose $a = 1$ and large oscillations to obtain a situation similar to that of Section XVII and we select the main trajectory $\phi_0 = 0$ which corresponds to the largest electronic current:

$$Y = \psi_0 - 0.337 \sin \psi_0 - 7.6 \sin \frac{\psi_0}{11.5} \cos (0.91\psi_0) \tag{166}$$

since $\phi_0 = 0$ yields $\phi = \Omega\psi_0$

hence

$$\frac{\psi_0}{2}(1 - \Omega) = \frac{\psi_0}{11.5} = 0.887\psi_0$$

$$\Phi = \frac{\phi}{2}\left(1 + \frac{1}{\Omega}\right) = \frac{\psi_0}{2}(1 + \Omega) = 0.91\psi_0$$

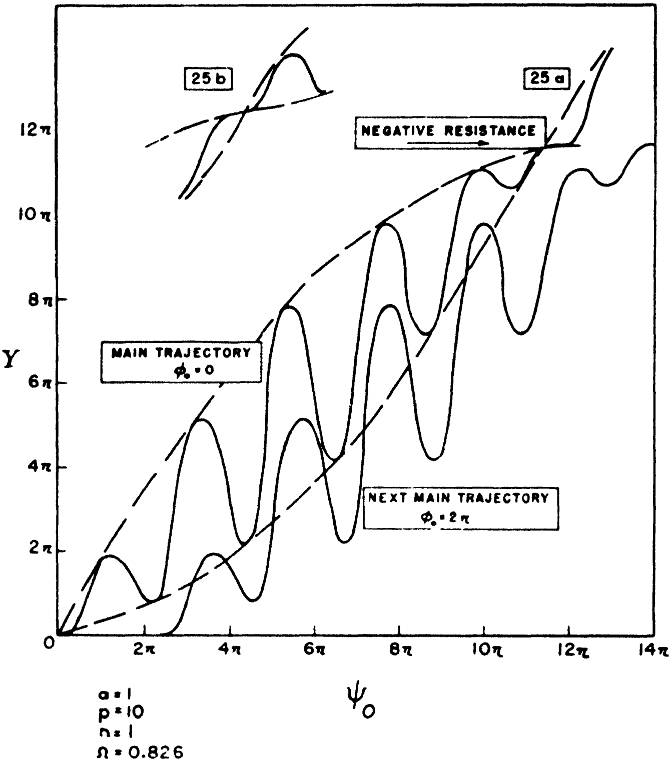


FIG. 25

This main trajectory is plotted on Fig. 25. Negative resistance is obtained for $\psi_0 = 11.5\pi$ which corresponds to *a node in the oscillations*. The exact shape of the trajectory near the node is different for p even (Fig. 25*a*) or odd (Fig. 25*b*). But in both cases the difference between ω and $2\omega_H$ results in beats between forced and free oscillations and condition (124) indicates a node in the beats. This means *no cavitation in* *front of the anode,* hence the disappearance of an effect which was obviously the cause of all our troubles in Section XVII. The beginning of the trajectory is very similar to that of Figs. 22 or 23 for exact resonance.

The next main trajectory ($\phi_0 = 2\pi$) was also drawn on Fig. 25 to visualize the bunching effect. It is parallel to the first one at a distance $2\pi/\Omega = 2.42\pi$. The points of negative resistance did not correspond to zero $\dot{y}$ velocity on the unperturbed trajectory (Fig. 24), and they do not coincide with these points on the perturbed trajectory either. Anyhow, the present discussion clearly shows why it was impossible to obtain negative resistance for exact resonance.

It must be noted also that the closest the resonance, the longest the transit time, since ω close to $2\omega_H$ means that it takes a long time to reach the first node of the beats.

The plane problem may certainly be considered as a special example of cylindrical structures, and it seems a safe guess that negative resistance should always be related with the appearance of a node in the beats. Conditions prevailing in cylindrical magnetrons differ, however, seriously from those found in the plane structure. A plane magnetron has just one well-defined proper frequency $2\omega_H$ whereas cylindrical structures have a whole frequency spectrum, starting from $2\omega_H$ near the cathode (radius a) and extending down to $\omega_H \sqrt{2 + 2(a^2/b^2)}$ near the cathode of radius b. This means that beats and nodes will appear in all cases and that there is no such thing as exact resonance. Negative resistance should probably correspond to a node near the anode

APPENDIX

Exact electromagnetic equations for a magnetron, with a discussion of some approximations usually introduced.

Let us rewrite the complete electromagnetic equations and check the kind of approximations involved in the theory developed in the main part of this paper. We start with the definition of the total current density

$$-\vec{J} = \rho\vec{V} + \epsilon_0 \frac{\partial}{\partial t}\vec{E} \tag{A.1}$$

The minus sign is chosen to yield a positive current in the direction of motion of electrons ($e < 0, \rho < 0$). Maxwell's equation directly result in

$$\operatorname{div} J = 0 \tag{A.2}$$

and read, with our sign convention

$$\operatorname{curl} H = -J \qquad \operatorname{div} H = 0 \tag{A.3}$$

$$\operatorname{curl} E = -\mu_0 \frac{\partial H}{\partial t} \qquad \operatorname{div} E = \frac{\rho}{\epsilon_0} \tag{A.4}$$

We assume a plane structure, with a cathode at $y = 0$, anode at $y = d$ and a magnetic field H_0 along z, to which we must add a variable part H_{1z}

$$H_z = H_0 + H_{1z} \qquad \omega_H = -\frac{1}{2}\mu_0 \frac{e}{m} H_0 \quad \text{Larmor} \tag{A.5}$$

Electronic motions are now given by equations of the type

$$\dot{x} = \frac{e}{m} E_x + \mu_0 \frac{e}{m}(\dot{y}H_z - \dot{z}H_y)$$

Singling out the H_0 terms, we obtain

$$\begin{aligned} \frac{d}{dt}(\dot{x} + 2\omega_H y) &= \frac{e}{m} E_x + \mu_0 \frac{e}{m}(\dot{y}H_{1z} - \dot{z}H_y) \\ \frac{d}{dt}(\dot{y} - 2\omega_H x) &= \frac{e}{m} E_y + \mu_0 \frac{e}{m}(\dot{z}H_y - \dot{x}H_{1z}) \\ \dot{z} &= \frac{e}{m} E_z + \mu_0 \frac{e}{m}(\dot{x}H_y - \dot{y}H_x) \end{aligned} \tag{A.6}$$

The z dependance of all quantities, and also the H_{1z} terms were ignored in the main paper. This may be justified in the following way.

Let us assume no current J_z to flow in the z direction, and $J_x J_y$ to depend only upon x,y,t but not upon z. Equation A.2 yields

$$\frac{\partial J_x}{\partial x} + \frac{\partial J_y}{\partial y} = 0 \tag{A.7}$$

and we take

$$\frac{\partial H_{1z}}{\partial x} = J_y \qquad \frac{\partial H_{1z}}{\partial y} = -J_x \tag{A.8}$$

which satisfies eq. A.3 provided

$$H_x = 0 \qquad H_y = 0 \tag{A.9}$$

and the remaining Maxwell's equations A.4 yield

$$\frac{\partial E_z}{\partial y} = \frac{\partial E_y}{\partial z} \qquad \frac{\partial E_x}{\partial z} = \frac{\partial E_z}{\partial x} \qquad \frac{\partial E_y}{\partial x} - \frac{\partial E_x}{\partial y} = -\mu_0 \frac{\partial H_{1z}}{\partial t}$$

Hence we may choose

$$E_z = 0 \qquad E_y, E_x \text{ functions of } x, y, t \text{ not } z \tag{A.10}$$

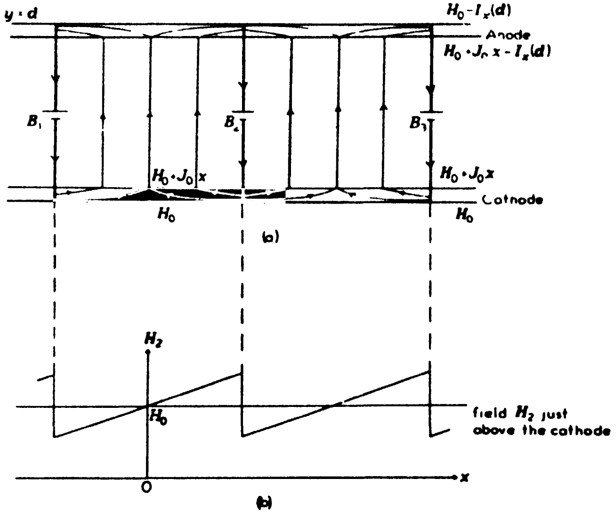


FIG. 26

Conditions (A.9) and (A.10) give no electronic motion and no displacement current along z, that is consistent with our initial assumption

$$J_z = 0 \tag{A.11}$$

and we are left with the equations of motion (A.6) with H_{1z} terms but no H_y terms (on account of A.9).

This rigorous method contains two new features, namely eq. A.8 and the H_{1z} terms in eq. A.6.

There is a curious peculiarity in this solution. Since we have a direct current between cathode and anode, J_y has a non-zero average $\bar{J}_y$, and consequently H_{1z} increases linearly with x.

In an actual structure this increase of H_{1z} could be avoided by a regular distribution of batteries $B_1B_2B_3$ inside the structure. The inverse currents flowing through these batteries compensate the direct current due to moving electrons and the resulting field distribution is shown on Fig. 26. A more detailed analysis of the field can be found in the report [11] In a cylindrical magnetron, one single battery inside the tube would suffice (Fig. 27).

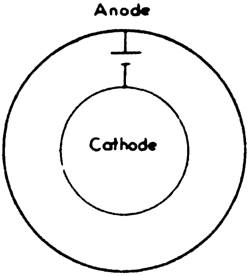


FIG. 27

The approximations involved in the main paper consist in neglecting the H_{1z} variable field due to electronic currents inside the magnetron. A numerical computation proves that this additional field always remains very small and can be neglected in front of the constant H_0 field.

REFERENCES

1. Moullin, E. B. *Proc. Cambridge Phil. Soc.*, **36**, 94 (1940).
2. Slater, J. C. Theory of the Magnetron Oscillator. M.I.T. Radiation Laboratory. Report V.5.S. August, 1941.
3. Page, Leigh, and Adams, N I, Jr., Space charge in plane magnetron. *Phys Rev.*, **69**, 492 (1945)
4. Llewellyn, F. B. Vacuum tube electronics *Proc Inst Radio Engrs.*, **21**, 1532 (1933)
   Fay, Samuel, and Shockley, W *Bell System Tech. J.*, **17**, 49 (1938).
5. Brillouin, L. Magnetron I. *Phys. Rev.*, **60**, 385 (1941); Magnetron II. *Phys. Rev.*, **62**, 166 (1943); Magnetron III. *Phys. Rev.*, **63**, 127 (1943), Influence of space charge. *Phys. Rev.*, **70**, 187 (1946).
6. Brillouin, L. *Proc. Inst. Radio Engrs.*, **32**, 216 (1944)
7. Brillouin, L. *Elec. Commun.* **22**, 110, 212 (1944 1945), **23**, 458 (1946).
8. Brillouin, L Electronic Theory of the Plane Magnetron, A.M.P. Report No. 129.1R, December, 1944.
9. British reports, C.V.D. Mag. No. 12, 16, 23, 30, by D. R. Hartree and E. C. Stoner.
10. Brillouin, L. Oscillations in a Plane One-Anode Magnetron, A.M.P. Report No. 129.3R, May 1945.
11. Brillouin, L. Exact electromagnetic equations for plane magnetron, A.M.P. Memo 129.1M, May, 1945.

Electronic Theory of the Cylindrical Magnetron

L. BRILLOUIN AND F. BLOCH

International Business Machines Corporation, New York, and Stanford University, Stanford, California

CONTENTS

Page
I. Summary and Introduction 145
II. Basic Assumptions for Steady Conditions 147
III The Problem's Equations Static Case 148
IV. Fundamental Equation of Motion 150
V. Discussion of the Fundamental Equation 151
VI. Small Current, Small Oscillations 153
VII. Mathematical and Graphical Discussion of the Self-Consistent Trajectories 155
VIII. Standard Static Characteristics of Cylindrical Magnetrons 159
IX. Physical Interpretation 163
  1. Magnetrons with $R < 2$ 164
  2. Magnetrons with $R > 2$ 165
X. Cylindrical Magnetron under Variable Conditions 166
XI. Discussion of the Solution Obtained for Small Current 168
XII. Limits of Validity of the Single Stream Solution 171
XIII. Small Oscillations in a Cylindrical Magnetron 174
XIV. Calculation of the Anode Voltage 176
XV. Resistance and Reactance of a Cylindrical Magnetron 179
References 181

I. SUMMARY AND INTRODUCTION*

The aim of this report is to present a summary of the theory for cylindrical magnetrons. Solutions of the general equation for "self-consistent trajectories" in the static case are obtained either by approximations or by numerical integration on the differential analyzer of the Massachusetts Institute of Technology.

Similar computations had been already performed by D. R. Hartree, W. P. Allis, or E. C. Stoner, but it was felt necessary to extend the field of integration beyond the limit of the first maximum, where all these authors stopped their computations. This continuation of the curves has no physical meaning in the steady case, but may be useful later for the discussion of oscillating magnetrons. Furthermore, these curves are very useful for a better understanding of the continuous distortion of magnetron characteristics when the anode radius is increased.

---

* This paper is based on work done by the authors for the Office of Scientific Research and Development under contracts with Columbia University (L. B.) and Harvard University (F. B.).

The *static characteristic of a magnetron* is an important information to be obtained from the theory, since it represents the practical summary needed by technicians. These characteristics are completely discussed in this report and some important results are underlined.

It is well known that: (I) plane magnetrons, (II) cylindrical magnetrons with $R < 2.273$ (R = ratio of anode radius to cathode radius), or (III) cylindrical structures with $R > 2.273$ should behave differently. For case I (plane) the static characteristic yielding current C with respect to voltage V, has been given in a preceding report. It is very interesting to note that the characteristics of cylindrical magnetrons $R < 2.273$ (case II) differ very little from the plane characteristics and that the difference can be ignored for most practical purposes. These characteristics explain the possibility for sustaining oscillations with a magnetron, since they exhibit some typical branches with *negative resistance.*

In case III ($R > 2.273$) a complete theoretical characteristic can still be drawn which is obtained by a continuous distortion from the preceding ones, and offers branches with negative resistance, but some special physical limitations appear in the static case and cause us to erase large parts of the curve, including the negative resistance branches. These limitations however cannot be maintained in case of oscillations, and the negative resistance branches reappear on this occasion.

The whole discussion is found in Sections VIII and IX, together with a typical characteristic for the case $R = 3.75$. Some details are worth mentioning.

(1) *The current at cutoff voltage* is, in all cases, almost three-fourths of the Langmuir current for a similar diode without magnetic field. The actual current ratio remains between 0.65 to 0.8 when R varies from 1 to 6.

This result had been obtained by J. C. Slater, using some formulas proposed by Allis. The proof was far from satisfactory, and the same result is now obtained directly from the general theory.

(2) In the plane problem, it was found that for large currents and high voltage, the magnetron's characteristics ran asymptotically to the Langmuir's curve, displaced by one-tenth of the cut-off voltage. In other words, if

$$V_L = KI^{2/3}$$

is Langmuir's formula for the diode, the magnetron's limit came out as

$$V \rightarrow \tfrac{1}{10}V_{co} + KI^{2/3}$$

A similar result is obtained for all cylindrical magnetrons on which computations have been made. The coefficient may slightly differ from $\frac{1}{10}$ but remains of that order of magnitude.

The most important difference between plane and cylindrical magnetrons should be found in their frequency spectrum: plane structures possess only one frequency $2\omega_H$, whereas cylindrical structures have a spectrum extending from $\omega_H \sqrt{2 + (2/R^3)}$ to $2\omega_H$ (ω_H, Larmor frequency). Sections X to XV are devoted to a discussion of a magnetron operated under variable conditions, and the first point investigated is the validity of the single stream solution. It is found that the discussion is much harder than for the plane structure. The first electrons emitted by the cathode will intercross at a distance $R = 1.434$, but their trajectories will disentangle themselves afterward, and steady double stream motion is impossible below $R = 2.273$. Other difficulties are found if there is no cause for energy dissipation in the circuit, since electronic oscillations in the space charge should intercross if they keep going for too long a time.

The conditions for negative resistance in a cylindrical one-anode magnetron have been obtained (Section XV) and have many points in common with the corresponding results for a plane magnetron.

It would be very interesting to check experimentally these theoretical results. They are rigorous and contain the fundamental assumption that end effects can be safely neglected. The theory applies only to single stream motion and indicates the limits of validity of such solutions. It remains an open question whether multiple stream motions actually play any important role in a single anode magnetron. In Sections X–XV it is further assumed that the current is small (eq. 49).

II. BASIC ASSUMPTIONS FOR STEADY CONDITIONS

The magnetron structure is supposed to contain a cylindrical cathode of radius a surrounded by a coaxial cylindrical anode of radius b with a constant magnetic field H in the direction of the axis of symmetry of the structure. The cathode is maintained at potential zero and emits electrons without velocity. The anode is at potential V and a steady current I flows through the tube. This total current may result from two currents flowing in opposite directions

$$\begin{aligned} I &= I_1 - I_2 \\ J &= I_1 + I_2 \geq |I| \end{aligned} \tag{1}$$

and the current entering the general equations is the sum J of the absolute values of both partial currents, rather than the actual total current I.

In the cylindrical magnetron, only two cases can be realized in a static case:

(1) *One virtual cathode* where all electrons stop and start backwards, this means

$$I_1 = I_2 \qquad J = 2I_1 \qquad I = 0 \tag{1a}$$

and yields a total current zero on the anode. The cathode emits a current I_1 and receives an equal back current (back bombardment).

(2) *No virtual cathode*

$$I_2 = 0 \qquad J = I_1 = I \tag{1b}$$

electrons never stop between cathode and anode. More than one virtual cathode cannot be obtained. At each point in the tube, the theory yields the absolute value $|v_r|$ of the v_r component of electrons' velocity; and the total space charge ρ. This total space charge may be divided into two parts ρ_1 and ρ_2, consisting of electrons with $\pm v_r$ velocities:

$$\begin{aligned} I_1 &= -2\pi r\rho_1|v_r| \\ I_2 &= +2\pi r\rho_2|v_r| \\ I &= -2\pi r(\rho_1 - \rho_2)|v_r| \\ J &= -2\pi r\rho|v_r| \end{aligned} \tag{2}$$

I and J are the currents per unit of length z of the structure. The signs are chosen so as to yield a positive current when electrons ($\rho < 0$) run from cathode to anode ($v_r > 0$). The absolute value $|v_r|$ of the radial velocity may become zero on the cylindrical surface

$$r = r_{p1} \cdots \tag{3}$$

if there is a "*virtual cathode*."

The fundamental equations will be summarized in the next section. They are practically the same as those used in an earlier paper.[1]

It will be shown that the solution proposed in this former paper corresponds to a limiting case and presents itself as the limit of curves whose structure is more elaborate.

III. THE PROBLEM'S EQUATIONS—STATIC CASE

Taking cylindrical coordinates r,θ,z (z being the axis of the structure), we assume electronic motions to take place in the r,θ plane only, as in the case for an infinitely long structure. The laws of motion are:

$$\left\{\begin{aligned} m\ddot{r} &= eE + \mu_0 eHr\dot{\theta} + mr\dot{\theta}^2 \\ m\frac{d}{dt}(r^2\dot{\theta}) &= -\mu_0 eHr\dot{r} \qquad MKS \text{ units} \end{aligned}\right. \tag{4}$$

E is the electric field, H the magnetic field, and the H term represents the Lorentz force, the $mr\dot{\theta}^2$ term is the usual "centrifugal force." The second equation expresses the law of conservation of angular momentum and can be readily integrated

$$\begin{aligned} \dot\theta &= \omega_H\left(1 - \frac{a^2}{r^2}\right) \\ \omega_H &= -\frac{1}{2}\mu_0\frac{e}{m}H \qquad \text{Larmor's rotation} \end{aligned} \tag{5}$$

Electrons are emitted without velocity on the cathode $r = a$, hence the value of the constant in eq. 5. Next to these equations we have the Poisson relation

$$\frac{1}{r}\frac{\partial}{\partial r}(rE) = \frac{\rho}{\epsilon_0} \tag{6}$$

and we derive the Llewellyn formula when we compute the d/dt derivative of (rE) along an electron trajectory

$$\frac{d}{dt}(rE) = r\frac{\partial E}{\partial t} + \dot r\frac{\partial}{\partial r}(rE) = r\left(\frac{\partial E}{\partial t} + \dot r\frac{\rho}{\epsilon_0}\right) \tag{7}$$

according to eq. 6. In a problem with varying currents $\epsilon_0\,\partial E/\partial t$ represents the displacement current density and $\rho\dot r$ the electronic current density, hence

$$\begin{aligned} \frac{d}{dt}(rE) &= -\frac{J}{2\pi\epsilon_0} \\ J &= -2\pi r\left(\epsilon_0\frac{\partial E}{\partial t} + \rho\dot r\right) \end{aligned} \tag{8}$$

the minus sign corresponds to the definition (eq. 2). The formula is thus proved for single stream electronic motion with variable J current. In case of a double stream static motion (eq. 2) it is also valid provided we follow the motion of an electron always moving forward, and take

$$\frac{\partial E}{\partial t} = 0 \qquad \dot r = |v_r| = 0$$

In electrostatic units $\epsilon_0 = \frac{1}{4\pi}$, the right-hand term in eq. 8 is $-2J$.[1]

Integrating eq. 8 we obtain

$$rE - aE_0 = -\frac{J}{2\pi\epsilon_0}\tau \tag{9}$$

where τ is the transit time for a forward moving electron and E_0 is the field on the cathode. If saturation is not reached (space charge limited current), we must assume $E_0 = 0$.

IV. FUNDAMENTAL EQUATION OF MOTION

The field distribution can be eliminated by combining eqs. 4, 5, and 9. We assume $E_0 = 0$ (no saturation) and obtain

$$\begin{aligned} \ddot{r} &= \frac{e}{m} E + r\dot{\theta}^2 - 2\omega_H r\dot{\theta} \\ \ddot{r} &= -\frac{e}{m}\frac{J}{2\pi\epsilon_0}\frac{\tau}{r} + \omega_H^2 r\left(\frac{a^4}{r^4} - 1\right) \end{aligned} \tag{10}$$

The problem, as a whole, is one of "self-consistent" field. Starting from the field distribution, trajectories are computed; then space charge is obtained which must yield the same field distribution from which the discussion was started. Equation 10 can be described as an equation for self-consistent trajectory in the steady case. It was first obtained by Hartree.

The case of the plane magnetron is obtained at the limit when

$$r = a + y, \quad y \ll a$$

The current J_{cyl} for the cylindrical problem is the current per unit of length while for the plane case we used J_p for the current density per unit square area, hence

$$J_{cyl} = 2\pi r J_p$$

and it is just a matter of expanding in powers of y/a to obtain the plane case.

It is convenient to introduce dimensionless ratios such as

$$R = r/a \tag{11}$$

which means that we use the cathode radius as unit of length, but

$$\left(-\frac{e}{m}\right)\frac{J}{4\pi\epsilon_0}\frac{1}{\omega_H^3} = L^2 \tag{12}$$

is the square of a characteristic length L ([1] eq. 33), which was found to play a prominent role in the behavior of the magnetron. Equation 10 thus reduces to

$$\ddot{R} = 2\left(\frac{L}{a}\right)^2 \omega_H^3 \frac{\tau}{R} - \omega_H^2\left(R - \frac{1}{R^3}\right) \tag{13}$$

Next we note that Larmor's angular velocity ω_H provides a natural time unit. Let us define

$$\theta = \omega_H \tau \tag{14}$$

we obtain

$$\frac{d^2R}{d\theta^2} = 2\left(\frac{L}{a}\right)^2 \frac{\theta}{R} - R + \frac{1}{R^3} \tag{15}$$

which is Allis'[3] eq. IV-67. Another way to write the same equation consists in taking

$$\phi = 2\left(\frac{L}{a}\right)^2 \theta = \frac{J\tau}{2\pi\epsilon_0\omega_H{}^2a^2}\left(-\frac{c}{m}\right) \tag{16}$$

which yields

$$\begin{aligned} \mu\frac{d^2R}{d\phi^2} &= \frac{\phi}{R} - R + \frac{1}{R^3} \\ \mu &= 4\frac{L^4}{a^4} = \left(\frac{e}{m}\right)^2 \frac{J^2}{4\pi^2\epsilon_0{}^2\omega_H{}^6a^4} \qquad \phi = \sqrt{\mu}\,\omega_H\tau \end{aligned} \tag{17}$$

This is the equation of a nonlinear oscillator, the equilibrium position of which is moving as a function of the reduced time ϕ, while the "mass" μ depends upon current and magnetic field.

The *instantaneous position of equilibrium* R_0 is defined as the point for which the "accleration" $d^2R/d\phi^2$ vanishes:

$$\phi = R_0{}^2 - \frac{1}{R_0{}^2}$$

hence

$$R_0{}^2(\phi) = \frac{\phi}{2} + \sqrt{(\phi^2/4) + 1} \tag{18}$$

The solution we need is

$$\frac{r}{a} = R = F(\phi) \tag{19}$$

which must satisfy eq. 17 with the *initial conditions*

$$\phi = 0, \qquad R = 1, \qquad \dot{R} = 0, \quad (\ddot{R} = 0) \tag{20}$$

since electrons leave the cathode ($r = a$, $R = 1$) at time $\tau = 0$ ($\phi = 0$), with zero velocity, and also zero acceleration.

V. DISCUSSION OF THE FUNDAMENTAL EQUATION

The voltage at any point is directly related with the kinetic energy of the electrons, since electrons leave the cathode ($V = 0$) with zero velocity:

$$v^2 = -\frac{2e}{m}V = \dot{r}^2 + \omega_H{}^2\left(r - \frac{a^2}{r}\right)^2$$

Introducing $R = r/a$ and ϕ from eq. 17, we obtain

$$\dot{r} = \frac{dr}{dt} = a\frac{dR}{d\phi}\frac{d\phi}{dt} = a\sqrt{\mu}\,\omega_H R' \text{ with } R' = \frac{dR}{d\phi}$$

and we may introduce a dimensionless expression for energy

$$-\frac{e}{ma^2\omega_H^2}V = \frac{1}{2}\mu R'^2 + \frac{1}{2}\left(R - \frac{1}{R}\right)^2 \qquad (21)$$

We shall first consider the mathematical problem and discuss it without referring to the physical problem. Later on, questions will be raised about the limits of validity of the equation itself, and the physical interpretation of the mathematical solutions will be obtained. Figure 1 is a graph of the R_0 curve (eq. 18) around which our solutions will oscillate. This curve is asymptotic to a parabola $R_0^2 = \phi$.

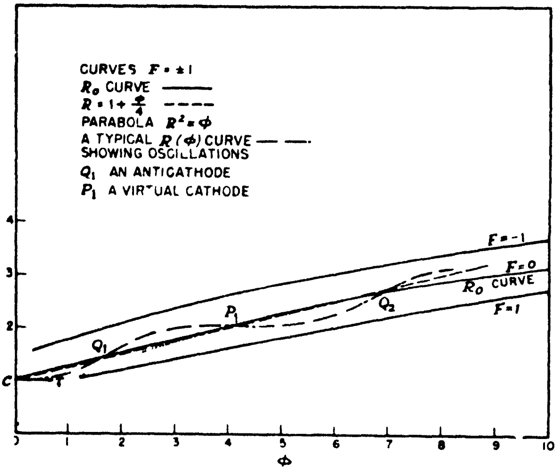


FIG. 1

The representative point has a zero acceleration when it crosses the R_0 curve (eq. 18) and its acceleration (eq. 17) always brings it back towards the R_0 curve, which results in oscillations. Curves corresponding to $F = \mu R'' = \pm 1$ are shown on the graph of Fig. 1.

Initial conditions are represented by the initial horizontal tangent CT which makes an angle $\frac{1}{4}$ with the R_0 curve. The circumstances are similar to those found for an oscillator with no damping, non-linear response, and an initial velocity. This comparison visualizes the oscillatory character of the motion.

The solution previously proposed[1] did not take into account these oscillations. A solution was worked out near the cathode, when the curve keeps near the original tangent CT, and it had been assumed that the parabola $R^2 = \phi$ represented a good approximation of the curve at large distance.

Attempts made at joining these two curves did not succeed, which is easy to understand, since the oscillatory character of the curve had been overlooked.

VI. SMALL CURRENT, SMALL OSCILLATIONS

The case of small currents, corresponding to $\mu < 1$ leads to curves exhibiting small oscillations around the $R_0(\phi)$ average curve defined in eq. 18.

Starting from eqs. 17 and 18

$$\mu R R'' = \phi - R^2 + \frac{1}{R^2}$$

we assume

$$R = R_0 + f \qquad |f| \ll R_0 \lesssim 1$$

It is easily proved by direct computation that R_0'' remains always small, its maximum value 0.063 being obtained for $\phi = 0$, $R_0 = 1$ The corresponding term can be neglected provided

$$f''| \gg |R_0''| \leq 0.063 \tag{22a}$$

Coming back to eq. 22, expanding, and dropping small terms, we find

$$\begin{aligned} \mu(R_0 f'') &= -2f(R_0 + R_0{}^{3}) \\ f'' + k^2 f &= 0 \\ \mu k^2 &= 2 + \frac{2}{R_0{}^4} \end{aligned} \tag{23}$$

where k is 2π times the frequency in ϕ units, hence the actual value

$$\omega = k\frac{d\phi}{dt} = k\sqrt{\mu}\,\omega_H = \omega_H\sqrt{2 + \frac{2}{R_0{}^4}} \tag{24}$$

This means that, for small oscillations, we have to deal with a harmonic oscillator, the frequency of which changes slowly with respect to time (or to the ϕ variable), since $R_0(\phi)$ is a function of ϕ directly.

Such problems are known as cases of adiabatic transformation of the oscillator.

A solution of eq. 23 may be found as an oscillation with varying amplitude, and we write it

$$f = A(\phi)\sin\psi(\phi) \tag{25}$$

At $\phi = 0$ we must satisfy the initial conditions (20)

$$\phi = 0 \begin{cases} R = 1 \quad \text{but} \quad R_0 = 1 \quad \text{hence } f = 0,\ \psi = 0 \\ R' = 0 \quad \text{hence} \quad f' = -R_0' = -\frac{1}{4} \end{cases} \tag{26}$$

From eq. 25 we compute

$$f'' = (A'' - A\psi'^2) \sin \psi + (2A'\psi' + A\psi'') \cos \psi$$

Here we make the *assumption* which corresponds to "very slow frequency variation," namely,

$$A'' \ll A\psi'^2 \tag{27}$$

Dropping A'' and substituting in eq. 23 we obtain

$$\begin{aligned} &\sin \psi \text{ terms: } \psi'^2 = k^2 \\ &\cos \psi \text{ terms: } 2A'\psi' + A\psi'' = 0 \end{aligned} \tag{28}$$

The first relation, together with the condition $\psi = 0$ for $\phi = 0$ yields

$$\psi = \int_0^\phi k\,d\phi \tag{29a}$$

while the second relation reads

$$\begin{aligned} 2\frac{A'}{A} &= -\frac{\psi''}{\psi'} \\ A = \frac{D}{\sqrt{\psi'}} &= \frac{D}{\sqrt{k}} \end{aligned} \tag{29b}$$

where D is a constant, hence the f function:

$$f = \frac{D}{\sqrt{k}} \sin \left(\int_0^\phi k\,d\phi \right) \tag{30}$$

We still have to satisfy the second initial condition (26):

$$\phi = 0 \qquad f' = \frac{D}{\sqrt{k_0}}\,k_0 = -\frac{1}{4} \qquad D = -\frac{1}{4\sqrt{k_0}} = -\frac{\mu^{1/4}}{4\sqrt{2}} \tag{31}$$

where k_0 is the value $2/\sqrt{\mu}$ of k for $\phi = 0$, $R_0 = 1$. The final result for small μ, small oscillations is

$$\begin{aligned} R = R_0 + f &= R_0(\phi) - \frac{\mu^{1/4}}{4\sqrt{2k}} \sin \psi \\ R_0^2 - \frac{1}{R_0^2} = \phi \qquad \psi &= \int_0^\phi k\,d\phi \end{aligned} \tag{32}$$

The solution is good so long as f remains small (eq. 22):

$$|f|^2 = \frac{\sqrt{\mu}}{32k} < \frac{\mu}{64} \ll 1 \quad \text{since} \quad k > \sqrt{\frac{2}{\mu}} \text{ (eq. 23)} \tag{33}$$

and it is easily proved that (eq. 22a) is also satisfied, as well as eq. 27.

VII. Mathematical and Graphical Discussion of the Self-Consistent Trajectories

Some physical limitations to the use of the solutions obtained from the general eqs. 17 and 22 will be discussed later, but we want first to examine the general type of the solutions. For very small μ, we obtained the solution (32). It consists of oscillations about the R_0 curve of Fig. 1, as schematically shown on Fig. 2, and the curve crosses the R_0 curve at smaller and smaller angles (angle ≥ 0). The process changes in the neighborhood of $R \approx 2.27$ when the R_0 curve bends down, while the amplitude A and frequency ω undergo very little change. The electron trajectory crosses the R_0 curve at almost constant angles.

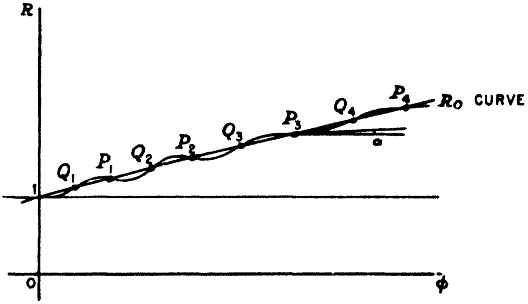


FIG. 2

These points of crossing are either anti-cathodes $Q_1Q_2Q_3 \cdots$ or quasi-virtual cathodes $P_1P_2P_3$. We are especially interested in the behavior of the trajectory at these quasi-virtual cathodes. For

$$1 < R < 2.27 \tag{34a}$$

the R' derivative at $Q_1Q_2 \cdots$ is positive, electrons are slowed down but not completely stopped at the quasi-virtual cathodes. Above that point

$$2.27 < R \tag{34b}$$

the derivative becomes negative, which means that electrons would reverse their motion. This is the point where a discussion of the physical interpretation of the mathematical results will be needed.

Let us first justify the figure 2.27 indicated above. It is based on the *assumption of very small oscillations*, $\mu \ll 1$.

The first horizontal tangent will be found on the R_0 line, hence for $\psi = 2m\pi$. From eq. 32 we obtain

$$R' = R_0' + \frac{\mu^{1/4}}{8\sqrt{2}\,k^{3/2}} \sin\psi - \frac{\mu^{1/4}}{4\sqrt{2}} \sqrt{k} \cos\psi$$

but $\psi = 2m\pi$ makes $\sin\psi = 0$ and $\cos\psi = 1$, hence the condition for horizontal tangent

$$R_0' = \frac{\mu^{1/4}}{4\sqrt{2}} \sqrt{k} \tag{35}$$

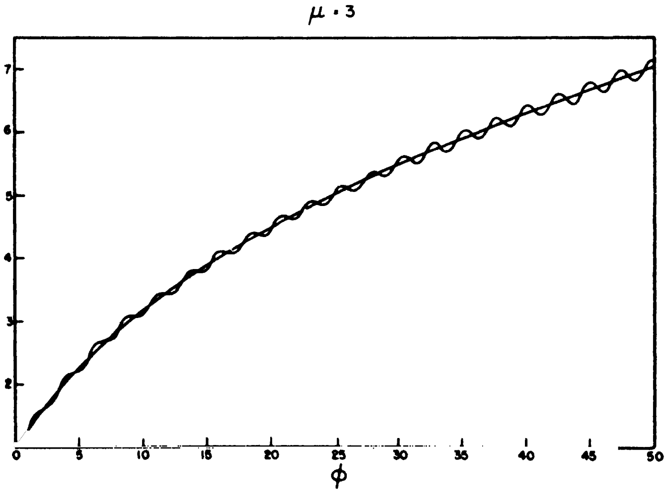


FIG. 3

Computing R_0' from eq. 18 and k from eq. 23 we find

$$\frac{2}{R_0(1 + R_0^{-4})} = \frac{1}{2^{1/4}} (1 + R_0^{-4})^{1/4}$$

hence

$$1 + R_0^{-4} = \frac{2}{R_0^{4/5}} \tag{36}$$

An obvious solution is $R_0 = 1$. The other solution lies below $2^{5/4} = 2.38$, and an exact computation gives 2.273. A number of solutions have been

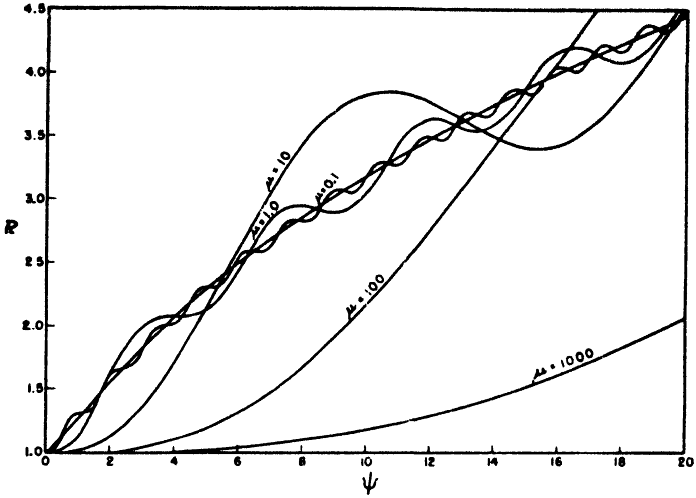


FIG. 4*a*

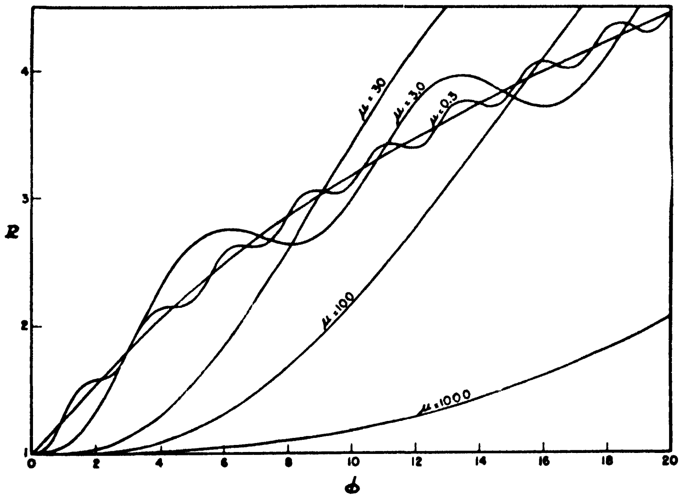


FIG. 4*b*

computed on the differential analyzer of the Massachusetts Institute of Technology and are shown on Figs. 3, 4*a*, 4*b*. Figure 3 is a graph of the solution for $\mu = 0.3$ and corresponds very clearly to the description given for very small μ_0. The first quasi-virtual *P* or anti-cathodes *Q* are found at

TABLE I

	Q_1	P_1	Q_2	P_2	Q_3	P_3	Q_4	P_4
$\psi =$	π	2π	3π	4π	5π	6π	7π	8π
$R =$	1.28	1.57	1.88	2.146	2.37	2.634	2.86	3.05
$R' = dR/d\phi$	0.53	0.05	0.47	0.01	0.42	−0.04	0.39	−0.08

The change from positive R' to negative R' at the quasi-virtual cathodes P takes place between P_2 ($R = 2.146$) and P_3 ($R = 2.634$) which are below and above our limit 2.273. Other curves, corresponding to

$$\mu = 0.1 \quad 0.3 \quad 1 \quad 3 \quad 10 \quad 30 \quad 100 \quad 1000$$

are shown on Figs. 4 and 5, and they all exhibit the same general behavior. The curve $\mu = 0.1$ has its first horizontal tangent at $R = 2.3$ just above the theoretical limit 2.27. As for the curve $\mu = 1$, it seems that its first horizontal tangent is at $R = 2.067$, which is not surprising since this μ value is too high for our approximations (36) to apply. A plot of the position $R = u$ of the first horizontal tangent (first real maximum, and actual virtual cathode) as a function of $b = \sqrt{\mu}$ was given by W. P. Allis, using Hartree's computations together with his own results. Our points are indicated by crosses (Fig. 5). The figure $n = 1.2 \cdots .10$ represents the number of quasi-virtual cathodes (or hesitations on the $R\phi$ curve) in front of the cathode (including the last virtual cathode). The curve jumps from n to $n + 1$ for certain μ (or b) values. The explanation will be better understood by looking at the curves of Fig. 4.

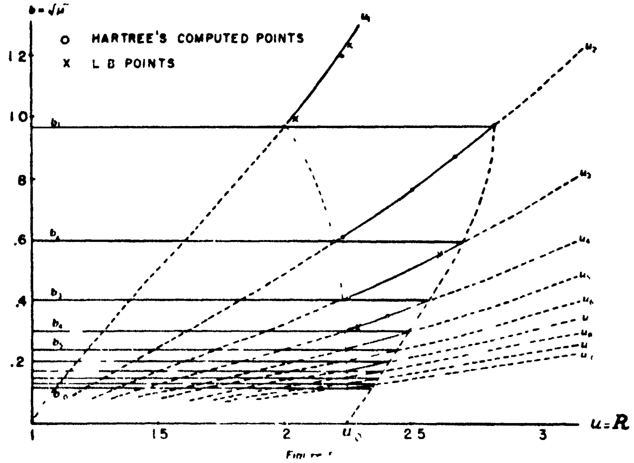


FIG. 5

For $\mu = 3$ the curve has a real maximum at $R = 2.761$, when μ decreases and reaches 1 we still find a maximum at $R = 2.067$, but it is almost a point of inflection. If μ decreases below 1, this point becomes an inflection with a positive derivative, and we have to go up to the next step to find a real maximum. This means that n jumps from 1 to 2, as shown on Fig. 5 (horizontal line $b = 0.97$).

Curves representing velocity as a function of the distance for plane magnetrons (preceding paper, Fig. 2*a*) were exact cycloids. A similar curve is plotted on top of Fig. 6, for the case $\mu = 1$. It shows the distortion of the original cycloid into a more complex curve. Knowing the velocity R' as a function of R, it is easy to compute the $\mu R'^2$ term and hence (eq. 21) the voltage as a function of R. The lower curve on Fig. 6 is a graph of $\mu R'^2$.

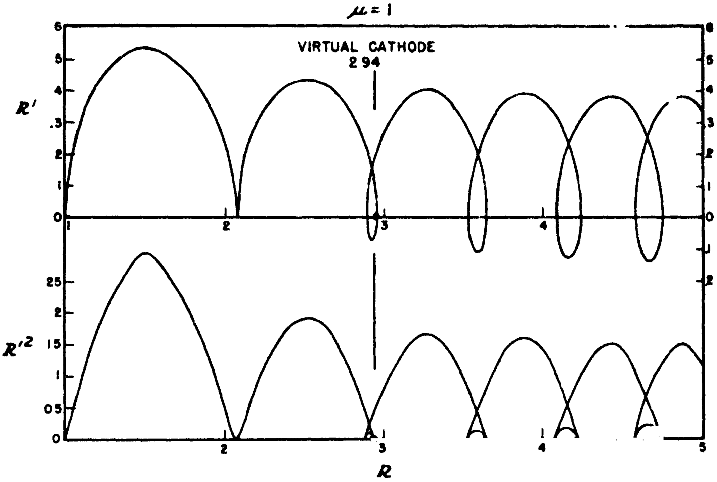


FIG. 6

VIII. STANDARD STATIC CHARACTERISTICS OF CYLINDRICAL MAGNETRONS

Standard current-voltage characteristics can be obtained for cylindrical magnetrons by the same method used for plane magnetrons. First the cutoff voltage V_{co} is chosen as unity and a nondimensional voltage measure U is defined

$$U = \frac{V}{V_{co}} = 1 + \mu \left[\frac{R'}{R - 1/R}\right]^2 \tag{37}$$

according to eq. 21. Next, we use the Langmuir curve for the diode without magnetic field

$$I_L = \left(-\frac{e}{m}\right)^{1/2} (2V)^{3/2} \frac{4\pi\epsilon_0}{9\beta^2 r} \qquad r = R \cdot a \tag{38}$$

where $\beta^2(R)$ is the function tabulated by Langmuir:

R	1	1.25	1.5	1.75	2	2.5	3	4	5	6	7	8	9	10	15	∞
β^2	0	0.045	0.116	0.2	0.275	0.405	0.512	0.665	0.775	0.818	0.867	0.902	0.925	0.94	0.978	1

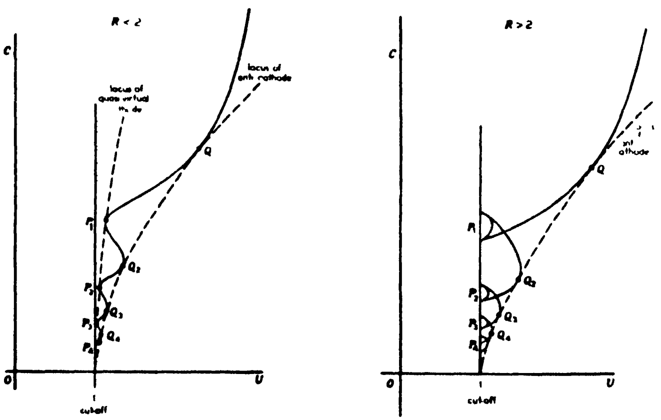


FIG. 7

We take as unit of current the Langmuir current I_{L0} at V_{c0} and define a nondimensional current measure:

$$C' = \frac{J}{I_{L0}} = \frac{9}{2}\beta^2 \frac{\sqrt{\mu}}{R^2[1 - (1/R^2)]^3} \tag{39}$$

Since eq. 21 yields

$$-\frac{e}{ma^2\omega_H^2} V_{c0} = \frac{1}{2}\left(R - \frac{1}{R}\right)^2 \qquad J = 0, \quad \mu = 0 \tag{40}$$

These definitions were introduced by J. C. Slater.

A standard characteristic curve for plane magnetron was given on Fig. 3 of a previous paper. Cylindrical magnetrons should exhibit curves of a different character, as shown on Fig. 7, where the details have been exaggerated. For $R < 2$ we noticed that the electrons never stop on the quasi-virtual cathodes, where they retain a small positive R'. This means that these quasi-virtual cathodes correspond to a voltage above cutoff (Fig. 7*a*). As a matter of fact, the residual voltage on these quasi-virtual cathodes is perfectly negligible. Let us take the example of Table I for the first quasi-virtual cathode P_1:

$$\mu = 0.3 \qquad R = 1.57 \qquad R' = 0.05 \tag{41}$$

$$\Delta U = \mu \left[\frac{R'}{R - 1/R} \right]^2 = 0.3 \frac{0.0025}{0.87} = 0.00086$$

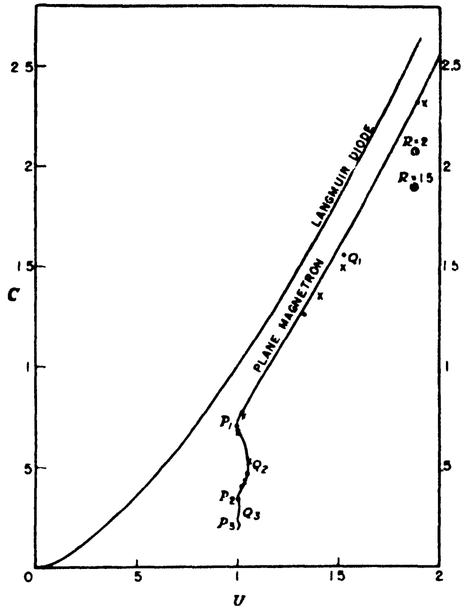


FIG. 8

The corresponding voltage is less than 0.1% above cutoff! And this is practically the largest ΔU which can be obtained.

It shall thus be predicted that the static characteristics of magnetrons of this type ($R < 2$) will differ very little from those of plane magnetrons.

This is shown on Fig. 8, where the standard characteristics of a plane magnetron is drawn, and a few points have been computed for $R = 1.5$ and $R = 2$. These points lie very near the curve for the plane magnetron.

When the ratio $R = r/a$ of anode-to-cathode radius is larger ($R > 2$), the behavior of the characteristics is materially modified, as shown on Fig. 7b. The rather strange shape of the curve near the vertical cut-off line comes from the similar parts of curves of Fig. 6b. An example of this type was (roughly) computed and plotted on Fig. 9. The question of the physical meaning of the different parts of the curve will be discussed later, but it was found advisable to draw the complete theoretical characteristics in order to show the continuous and progressive change in the shape of the curves corresponding to increasing R values.

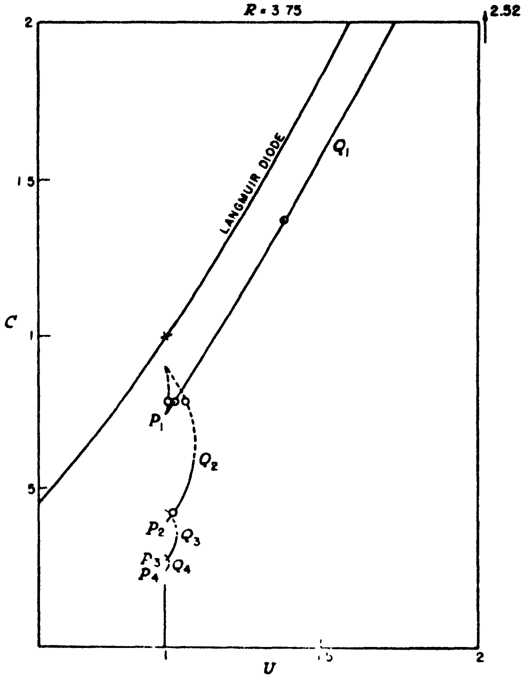


FIG. 9

For the plane magnetron, the first anti-cathode and virtual cathode were found at

$$\begin{array}{lll} \text{A.C.}\ (Q_1) & C = 1.44 \\ \text{V.C.}\ (P_1) & C = 0.72 \end{array} \qquad (42)$$

On the characteristics of Fig. 9, the *virtual cathode* is replaced by a triangular design, in which we choose the lower corner P_1 as most representative, since it represents the lower current obtained by decreasing slowly voltage and current from large values down to the cutoff. The point P_1 corresponds to the first horizontal tangent found on a μ curve (Figs. 4 or 5) when climbing the curve up from the origin.

The first *anti-cathode* is defined as the first point where a μ curve crosses the R_0 curve and obtains a point of inflexion. From the curves of Fig. 4 one may compute the positions of the P_1Q_1 points for various μ

and plot the curves of Fig. 10. The dashed curves result from the approximate solutions of Section VI for small μ. A very remarkable result is the stability of point P_1:

$$\text{First virtual cathode } P_1 \quad C \approx 0.75 \text{ for any } R \tag{43}$$

This result had been obtained by J. C. Slater from some general relations given by Allis, and it is now proved to represent a very good approximation.

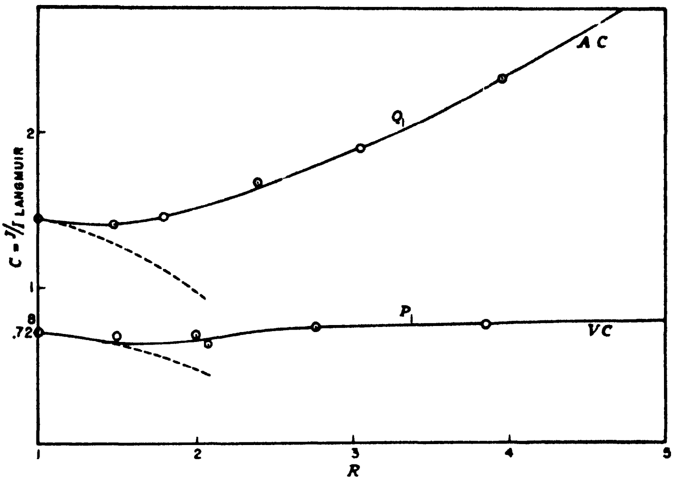


Fig 10

IX. Physical Interpretation

The preceding mathematical discussion was simply based on eq. 22 for self-consistent trajectories. From the physical point of view, there is an additional condition to be fulfilled, according to the discussion of Section III; at each point, we must obtain *only one value* for the *voltage* and *two opposite values for the radial velocity*. This means that the trajectories of Figs. 3 and 4 can be followed only up to their first horizontal tangent (virtual cathode). From this virtual cathode on, electrons do not proceed along the theoretical curve but start moving back to the cathode along the same trajectory which they have been following up from the cathode.

Looking at the curves of Fig. 6, they must be utilized up to the first virtual cathode ($\mu = 1$, $R = 2.94$) and the following loops of the curve cannot be reached. The triangular parts of curve, on lower Fig. 6 have no physical meaning.

In this respect, two different cases shall be distinguished:

1. *Magnetrons with $R < 2$*

R means the ratio of anode radius to cathode radius. For such structures there is no curve exhibiting any horizontal tangent. We proved in Section VII that the curves obtained for small μ values do not show any horizontal tangent below $R = 2.273$. For higher μ values, the limit is lowered and we noticed a horizontal tangent at $R = 2.067$ for $\mu = 1$. The overall limit should be practically near $R = 2$ (see Fig. 5).

Magnetrons with $R < 2$ cannot exhibit any virtual cathode, and their theoretical characteristics have the shape shown on Fig. 7*a*. Any arbitrary current value C can be used and the characteristic is perfectly continuous. The only solution with *zero total current* is the B solution

$$I_1 = I_2 = J = 0 \qquad \mu = 0$$

for which the electron trajectory, in $R\phi$ variables, is the R_0 curve itself. Our formulas of Section VI proved that for small μ the oscillations around the R_0 curve become smaller and smaller, until for $\mu = 0$ the trajectory reduces to the R_0 curve

This, at least, is the description in $R\phi$ coordinates, but if we revert to the usual variables, $R\tau$, we obtain a different description, since, according to eqs. 17 and 19

$$\mu = \left(\frac{e}{m}\,\frac{J}{2\pi\epsilon_0\omega H^3a^2}\right)^2 \qquad \phi = \sqrt{\mu}\,\omega_H\tau \tag{44}$$

When $J \to 0$, $\mu \to 0$, a *finite reduced time* ϕ means an *infinite transit time* τ, hence it takes an infinite time for an electron to reach a finite distance R. This peculiarity of the B solution was already emphasized by in previous papers.[1] [2]

The *B solution* actually consists of electrons running on circular orbits, with no radial velocity ($\dot{R} = 0$) and an angular velocity (eq. 5)

$$\dot{\theta} = \omega_H\left(1 - \frac{1}{R^2}\right) \qquad R = \frac{r}{a}$$

with a space charge density

$$\rho_0 = \epsilon_0\,\frac{m\omega_H^2}{2\pi e}\left(1 + \frac{1}{R^4}\right) \tag{45}$$

Speaking of the *static characteristics* of these magnetrons ($R < 2$) we found that it differs very little from the characteristics of plane magne-

trons (Fig. 8), and the P_1Q_2 branch exhibits *negative resistance* as in the plane case.

The only marked difference between these structures and the plane ones is found in their *spectrum of internal proper vibrations* (eq. 24) which extends from $2\omega_H$ to $\omega_H \sqrt{2 + 2}\ R^4$ while the plane magnetron exhibits only *one frequency* $2\omega_H$. This difference in internal vibration spectra corresponds to a different behavior of the magnetrons at high frequencies, but it does not appear on the static characteristics.

2. Magnetrons with $R > 2$

These structures may exhibit virtual cathodes, and their characteristics have the shape shown on Fig. 9. Let us discuss it more carefully, keeping in mind the fact that the curves of Fig. 4 must be followed only up to their first maximum. Let us take the specific example of Fig. 9, $R = 3.75$. For large currents (large μ) there is no maximum on the curves between $R = 1$ and 3.75, hence no special restriction. The curve $C = 0.79$ ($\mu = 10$) can be used only up to its maximum at $R = 3.85$, but the other points should not be used. This means that the whole Q_1P_1 branch can be retained, but that the small dotted triangle above P_1 is fictitious. Current values just below 0.79 ($\mu < 10$) do not correspond to any physical solution, since these trajectories do not reach the anode. Decreasing the current to about 0.6 ($\mu \sim 5$) we finally reach a trajectory with a horizontal tangent and inflexion just on the R_0 line (below $R = 3.75$) and its first maximum above 3.75. This trajectory gives a point in the neighborhood of Q_2 on the characteristics, and corresponds to an acceptable physical solution. Lower currents ($0.4 < C < 0.6$) and lower μ values ($0.25 < \mu < 0.5$ approximately) will then yield another branch Q_2P_2 of physical significance, and so on. On the curve of Fig. 9, the parts without physical meaning are drawn as dashed lines, parts with a physical meaning are drawn with a solid line. It is very significant that all the P_1Q_2 or P_2Q_3 *branches with negative resistance lose their physical meaning.* Such magnetrons with $R > 2$ should not be able to sustain oscillations at very low frequencies. But it seems a safe guess that the behavior of these magnetrons at high frequency may not differ too much from the behavior of magnetrons with $R < 2$. The negative resistance branch P_1Q_2 of Fig. 8 will be very much distorted at high frequencies, and the corresponding branch of Fig. 9 will certainly reappear. This seems especially plausible if one remembers that the condition forbidding loops in the trajectories is the static condition of energy conservation and definition of a potential. In nonstatic conditions with oscillating fields, there is no potential to be defined, loops are no more forbidden, and the branches P_1Q_2 or P_2Q_3 will reappear, with some distortion.

We may now *summarize* the discussion and state the possibilities for magnetrons operated *below cutoff* when the total *current I* is *zero:*

Plane magnetron both cases A and B, but only case B is stable (see preceding report).

Cylindrical magnetron $R < 2$ only case B; no possibility for a case A, since no virtual cathode can be obtained.

Cylindrical magnetron $R > 2$ cases A and B are both theoretically possible.

X. CYLINDRICAL MAGNETRON UNDER VARIABLE CONDITIONS

The general equation given in Section III can be easily rewritten for the case of a variable current I. We start from eqs. 8, 10, and 13 and we obtain

$$R\ddot{R} + \omega_H^2(R^2 - R^{-2}) = \frac{e}{ma} RE = 2\omega_H^2\varphi(t,t_0) \tag{46}$$

with $R = r\ a$

$$\varphi(t,t_0) = \frac{e}{ma^2 2\omega_H^2} rE = \frac{(-e)}{4\pi\epsilon_0 ma^2\omega_H^2}\int_{t_0}^{t} I(t')dt' \tag{47}$$

Electrons leave the cathode at t_0 with zero velocity and zero acceleration

$$t = t_0 \qquad R = 1 \qquad R = 0 \qquad \ddot{R} = 0 \tag{48}$$

the last condition means space charge limited current (no saturation) and is included in eq. 47, that makes $E = 0$ at t_0. Equation 46 yields the *self-consistent trajectories* $R(t_0,t)$. Approximate solutions can be found for small currents when

$$\epsilon(t) = \frac{1}{2\omega_H}\frac{\partial\varphi}{\partial t} = \frac{(-e)I(t)}{8\pi\epsilon_0 ma^2\omega_H^3} \ll 1 \tag{49}$$

that means that φ is a slowly varying function and changes very little during a Larmor period. Numerical computation shows that this condition is always satisfied in practice. We may now expand

$$R = R_0 + R_1 + R_2 \cdots \tag{50}$$

and obtain the successive approximations

$$\begin{aligned} &R_0^2 - R_0^{-2} = 2\varphi(t,t_0) \\ &\ddot{R}_1 + 2\omega_H^2 R_1(1 + R_0^{-4}) = -\ddot{R}_0 \\ &\ddot{R}_2 + 2\omega_H^2 R_2(1 + R_0^{-4}) = \omega_H^2(R_0^{-1} + 5R_0^{-5})R_1^2 \cdots \end{aligned} \tag{51}$$

The dominant term R_0 is given by

$$R_0 = \sqrt{\varphi + \sqrt{\varphi^2 + 1}} \tag{52}$$

and is a slowly varying function of time. If the current I is a small positive quantity, both φ and R_0 increase slowly, and the next approximation R_1 is a small oscillation. Let us introduce the frequency ω_1

$$\begin{aligned} \omega_1(t,t_0) &= \omega_H \sqrt{2 + 2R_0^4} \\ \ddot{R}_1 + \omega_1^2 R_1 &= -\ddot{R}_0 \end{aligned} \tag{53}$$

The second approximation represents the equation of motion of a harmonic oscillator, with slowly varying frequency ω_1, under an external force $-\ddot{R}_0$. We start from the homogeneous equation

$$\ddot{\xi} + \omega_1^2 \xi = 0 \tag{54}$$

and assume that we have obtained two independant solutions ξ_1, and ξ_2, normalized so as to give, at any time

$$\dot{\xi}_1 \xi_2 - \dot{\xi}_2 \xi_1 = 1 \tag{55}$$

The solution of eq. 53 can then be written in the following form

$$R_1(t,t_0) = \xi_1(t) \int_{t_0}^{t} \dot{R}_0(t')\xi_2(t')dt' - \xi_2(t) \int_{t_0}^{t} \dot{R}_0(t')\xi_1(t')dt' \tag{56}$$

that also satisfies the initial conditions (48). The next approximation R_2 can be obtained by a similar procedure.

The problem is to build the solution ξ_1 and ξ_2 of the homogeneous eq. 54, when the function ω_1 of eq. 53 is a slowly varying function.

We introduce a dimensionless variable θ that measures 2π times the number of cycles between t_0 and t

$$\frac{d\theta}{dt} = \omega_1 \qquad \theta(t,t_0) = \int_{t_0}^{t} \omega_1(t',t_0)dt' \tag{57}$$

and a new quantity

$$l = \log \sqrt{\omega_1} \qquad l' = \frac{dl}{d\theta} = \frac{1}{2\omega_1}\frac{d\omega_1}{d\theta} = \frac{1}{2\omega_1^2}\frac{d\omega_1}{dt} \ll 1 \tag{58}$$

according to eq. 49. With these variables, eq. 54 takes the following shape

$$\xi'' + 2l'\xi' + \xi = 0, \quad \xi' = d\xi/d\theta \tag{59}$$

A solution is found by taking

$$\xi = Ae^{-l+a+ib} \tag{60}$$

with a constant A and two real functions a and b:

$$b' = e^{-2a}$$
$$a'' + 4a = l'' + a'^2 + l'^2 + (e^{-4a} + 4a - 1) \qquad (61)$$

This has been written in such a way as to make all terms on the right negligible in case of small currents. Terms on the left are linear in the current and yield a solution

$$a(\theta) = \int_0^\theta l'(\theta_1) \cos 2(\theta - \theta_1) d\theta_1 \qquad (62)$$

Quadratic terms and higher order terms in the current may be neglected, but for some exceptional cases like l' varying in $\cos 2\theta$ or $\sin 2\theta$, that lead to a large increase of a. This special case of resonance will not be considered here. When this special case is excluded, eq. 62 represents a reasonable approximation, and we may consistently drop higher order terms in the expression of b. Starting from eq. 61 we expand

$$b' = 1 - 2a \cdots \qquad b = \theta + b_1 \cdots \qquad (63)$$

with $\qquad b_1 = -2 \int_0^\theta a(\theta_1) d\theta_1 = -\int_0^\theta l'(\theta_1) \sin 2(\theta - \theta_1) d\theta_1$

The terms kept in eq 63 are linear in the current and small compared to unity on account of eq. 58 The same is true for a. Combining eqs. 60, 62, and 63, we obtain the two independent solutions we need

$$\xi_1 = \frac{A}{\sqrt{\omega_1}} e^{a + i(\theta + b_1)} \qquad \xi_2 = \frac{A}{\sqrt{\omega_1}} e^{a - i(\theta + b_1)} \qquad (64)$$

since $e^{-l} = \frac{1}{\sqrt{\omega_1}}$ from eq 58

Elementary computations show that the normalization condition (55) is satisfied when

$$A = \frac{1}{\sqrt{2i}} \qquad (65)$$

since

$$\dot{\xi}_1 \xi_2 - \dot{\xi}_2 \xi_1 = \omega_1(\xi_1' \xi_2 - \xi_2' \xi_1) = \omega_1 A^2 2ib' e^{-2l+2a} = 2iA^2$$

on account of eqs 58 and 61. Equations 56 and 62 to 65 yield the formal solution of the problem for small currents.

XI. DISCUSSION OF THE SOLUTION OBTAINED FOR SMALL CURRENT

In using these equations we shall consistently keep only linear terms in the current. The selection of terms to be kept in the formulas is

based on the fact that $\dot{\varphi}$, R_0, a, b_1, are linear in the current, according to eqs. 47, 52, 58, 62, 63:

$$\dot{R}_C(t) = \frac{\partial\varphi}{\partial t}\frac{1}{R_0 + R_0^{-3}} = \frac{2\omega_H\epsilon(t)}{R_0 + R_0^{-3}} \tag{66}$$

on account of eq. 49. Hence

$$\dot{R}_0(t_0) = \omega_H\epsilon(t_0) \tag{67}$$

since $R_0(t_0) = 1$. Before using eq. 56 we integrate by parts:

$$R_1(t) = -\dot{R}_0(t_0)[\xi_1(t)\xi_2(t_0) - \xi_2(t)\xi_1(t_0)] - \xi_1(t)\int_{t_0}^{t}\ddot{R}_0(t')\xi_2(t')dt' + \xi_2(t)\int_{t_0}^{t}\ddot{R}_0(t')\xi_1(t')dt' \tag{68}$$

Now, for $t = t_0$, $\theta = 0$ we have $\omega_1 = 2\omega_H$ (eq. 53) and the exponents in eq. 64 vanish, hence

$$\xi_1(t_0) = \xi_2(t_0) = \frac{1}{2\sqrt{i\omega_H}} \tag{69}$$

Since we keep only linear terms in the current, we may drop correction terms in a and b_1 in eq. 64 and take

$$\xi_1(t) = \frac{1}{\sqrt{2i\omega_1(t)}}e^{-i\theta} \qquad \xi_2(t) = \frac{1}{\sqrt{2i\omega_1(t)}}e^{i\theta} \tag{70}$$

and eq. 68 becomes

$$R_1(t) = -\frac{\epsilon(t_0)}{2}\sqrt{\frac{2\omega_H}{\omega_1(t)}}\sin\theta - \int_{t_0}^{t}\frac{\dot{R}_0(t')}{\sqrt{\omega_1(t)\omega_1(t')}}\sin(\theta - \theta')dt' \quad \theta' = \theta(t') \tag{71}$$

Finally, we obtain the complete first order solution

$$R(t) = R_0(t) + R_1(t) \tag{72}$$

where, to summarize the definitions

$$\xi(t) = \frac{(-e)I(t)}{8\pi\epsilon_0 ma^2\omega_H^3} \ll 1 \tag{49}$$

$$\varphi(t) = 2\omega_H\int_{t_0}^{t}\epsilon(t')dt' \tag{47}$$

$$R_0(t) = \sqrt{\varphi + \sqrt{\varphi^2 + 1}} \tag{52}$$

$$\omega_1(t) = \omega_H\sqrt{2 + 2R_0^{-4}} \tag{53}$$

$$\theta(t) = \int_{t_0}^{t}\omega_1(t')dt' \tag{57}$$

$$\theta - \theta' = \theta(t) - \theta(t') = \int_{t'}^{t}\omega_1(t'')dt'' \tag{73}$$

As stated before, the solution thus obtained for small currents represents a slow radial drift (when $I > 0$) represented by R_0 and superimposed small oscillations with variable frequency.

A general theorem was proved for the plane magnetron: when the tube is operated below saturation (space charge limited current) and no negative currents are allowed, electron trajectories never cross each other and the motion retains the single stream character. Electrons that have been emitted from the cathode at a later time t_0 will always be found closer to it than those emitted at an earlier time, hence

$$\left(\frac{\partial R}{\partial t_0}\right)_{t\ \text{const}} < 0 \tag{74}$$

We must now discuss whether such a condition is satisfied for a cylindrical magnetron. Retaining only terms linear in the current we obtain from eqs. 71 and 72

$$\begin{aligned}\frac{\partial R}{\partial t_0} = \frac{\partial R_0}{\partial t_0} - \frac{1}{2}\frac{d\epsilon(t_0)}{dt_0}\sqrt{\frac{2\omega_H}{\omega_1(t)}}\sin\theta + \frac{\omega_2(t)}{2}\,\epsilon(t_0)\sqrt{\frac{2\omega_H}{\omega_1(t)}}\cos\theta \\ + \frac{\dot R_0(t_0)}{\sqrt{\omega_1(t)\omega_1(t_0)}}\sin\theta + \int_{t_0}^{t}\frac{\ddot R_0(t')}{\sqrt{\omega_1(t)\omega_1(t')}}[\omega_2(t) - \omega_2(t')]\cos(\theta - \theta')dt'\end{aligned} \tag{75}$$

Terms containing the derivatives $\partial\omega_1\ \partial t_0$ and $\partial R_0/\partial t_0$ have been neglected since they contribute only in higher orders, and we have introduced

$$\omega_2(t) = -\frac{\partial\theta(t)}{\partial t_0} = 2\omega_H - \int_{t_0}^{t}\frac{\partial\omega_1(t')}{\partial t_0}dt' \tag{76}$$

The integral in eq. 76 is linear in the current, but it increases indefinitely with the time, hence terms in ω_2 must be kept although they appear only multiplied with other factors that are already linear in the current. Furthermore

$$\frac{\partial R_0}{\partial t_0} = \frac{dR_0}{d\varphi}\frac{\partial\varphi}{\partial t_0} = -2\omega_H\frac{\epsilon(t_0)}{R_0 + R_0{}^{3}} \tag{77}$$

According to eq. 67 we have

$$\ddot R_0(t_0) = \omega_H\frac{\partial\epsilon(t_0)}{\partial t_0} \qquad \omega_1(t_0) = 2\omega_H$$

that shows that the second and forth terms in eq. 75 cancel each other, and we are left with

$$\begin{aligned}\frac{\partial R}{\partial t_0} = -\omega_H\epsilon(t_0)\left[\frac{2}{R_0 + R_0^{-3}} - \frac{\omega_2(t)\cos\theta}{\sqrt{2\omega_H\omega_1(t)}}\right] + \int_{t_0}^{t}\frac{\ddot R_0(t)}{\sqrt{\omega_1(t)\omega_1(t')}} \\ \times[\omega_2(t) - \omega_2(t')]\cos(\theta - \theta')dt'\end{aligned} \tag{78}$$

This is the expression that we have to discuss in order to find out whether it remains negative or not, and to check when the condition (74) for single stream holds.

The special case of a *plane structure* that was previously discussed is obtained when

$$r = a + y, \quad R = 1 + \frac{y}{a}, \quad \frac{y}{a} \ll 1 \quad \text{hence } R_0 = 1 \tag{79}$$

and

$$\omega_1 = 2\omega_H = \omega_2, \quad \theta = 2\omega_H(t - t_0)$$

The last integral in eq. 78 vanishes and

$$\frac{\partial y}{\partial t_0} = a\frac{\partial R}{\partial t_0} = a\omega_H\epsilon(t_0)[-1 + \cos\theta] < 0 \tag{80}$$

and condition (74) holds.

The characteristic feature of the plane problem is that there is only one constant proper frequency $2\omega_H$ instead of the variable frequencies ω_1 and ω_2 of the cylindrical problem. This remark will be very helpful in the comparison of both cases.

XII. Limits of Validity of the Single Stream Solution

In the *cylindrical case* the discussion is simplified when the current varies very slowly since R_0 is proportional to the current (eqs. 67 and 49) and $\dot{R}_0$ can be neglected for *slow varying currents.* Expression (78) reduces to the first bracket, that remains negative provided

$$\left|\frac{\omega_2}{\sqrt{2\omega_H\omega_1}}\right| < \frac{2}{R_0 + R_0^{-3}}$$

or, using eq. 53

$$\left|\frac{\omega_2}{\omega_1}\right| < \frac{1}{R_0}\left(\frac{2}{1 + R_0^{-4}}\right)^{3/4} \tag{81}$$

But eq. 76 shows that

$$\frac{\partial\omega_2}{\partial t} = -\frac{\partial\omega_1}{\partial t_0} = \frac{\partial\omega_1}{\partial t}\frac{I(t_0)}{I(t)} \tag{82}$$

since eqs. 47, 52, and 53 yield

$$\frac{\dfrac{\partial\omega_1}{\partial t_0}}{\dfrac{\partial\omega_1}{\partial t}} = \frac{\dfrac{\partial R_0}{\partial t_0}}{\dfrac{\partial R_0}{\partial t}} = \frac{\dfrac{\partial\varphi}{\partial t_0}}{\dfrac{\partial\varphi}{\partial t}} = -\frac{\epsilon(t_0)}{\epsilon(t)} = -\frac{I(t_0)}{I(t)}$$

On the other hand, Eqs. 53 and 66 give

$$\frac{\partial\omega_1}{\partial t} = -2\sqrt{2}\,\omega_H\frac{R_0}{R_0^5\sqrt{1 + R_0^{-4}}} = -4\omega_H^2 \times \frac{\sqrt{2}\,\epsilon(t)}{(R_0^4 + 1)^{3/2}} \tag{83}$$

hence

$$\frac{\partial\omega_2}{\partial t} = -4\omega_H{}^2 \frac{\sqrt{2}\,\epsilon(t_0)}{(R_0{}^4 + 1)^{5/4}} \tag{84}$$

Both ω_1 and ω_2 start from the common value $2\omega_H$ at t_0 and decrease progressively, since we assumed the current never to become negative, hence ϵ to remain always positive. For *constant current* we have obviously $\omega_2 = \omega_1$ and condition (81) yields

$$1 < \frac{1}{R_0}\left(\frac{2}{1 + R_0{}^{-4}}\right)^{5/4}$$
$$R_0 < 2.273 \tag{85}$$

that is the condition already obtained in Section VII, eqs. 34 and 36. For *slowly increasing current* $I(t) > I(t_0)$ for $t > t_0$ and ω_2 decreases more slowly than ω_1

$$\omega_1 \leq \omega_2 \leq 2\omega_H \tag{86}$$

ω_2 may even remain constant and equal to $2\omega_H$ when $I(t_0) = 0$ that corresponds to the first electron leaving the cathode at the start of magnetron operation.

The lower limit in eq 86 corresponds to the case of constant current and condition (85). The upper limit in eq. 86 leads to the condition

$$\frac{1}{R_0}\left(\frac{2}{1 + R_0{}^{-4}}\right)^{3/4} = 1 \tag{87}$$
$$R_0 = 1.134$$

For *slowly decreasing current*

$$\frac{I(t_0)}{I(t)} > 1 \quad \text{for } t > t_0$$

and according to eq. 82, ω_2 decreases faster than ω_1 hence

$$\omega_2 < \omega_1$$

since both quantities start from the same value $2\omega_H$ at t_0. The frequency ω_1 always remains positive, and we first discuss the case when ω_2 is positive too, in which case eq. 81 indicates that the single stream solution will retain its validity for R_0 values larger than 2.273 (eq. 83).

There is, however, a possibility that ω_2 becomes negative, and this will be better discussed on a typical example. Let us consider the usual process of "turning on" a magnetron by letting a small current flow from $t = 0$ to $t = T$, and keeping the current zero for $t > T$. The first electrons leaving the cathode at $t = 0$ will loose the single stream character when they reach a distance $R_d = 1.434$, but this must be a transitory trouble, since there cannot exist a stable double stream motion for distances smaller than 2.273. During the period of decreasing current, the critical distance becomes even larger. But what happens during the later time interval, when $t > T$ and $I = 0$?

Electrons run on successive concentric layers, corresponding to constant R_0 and ω_1 values for each of them. But ω_2 goes on decreasing indefinitely, according to eq. 84, hence there will come a moment when condition (81) cannot hold any more, since $\omega_2 \ \omega_{10}$ may increase above any limit. Single stream solutions will cease to exist at that time. This can be easily visualized; when the current is zero, electrons run on the average, on concentric layers and have small oscillations about their average circular trajectories. Electron at R_0 has a frequency $\omega_1(R_0)$ but his neighbor at R_0' has a different frequency $\omega_1(R_0')$. Their trajectories may run parallel to each other at the beginning, but this cannot last long, and after a time ($\frac{1}{2}$ beat period), they will oscillate with opposite phases and their trajectories will cross each other to disentangle themselves after a while. Such a situation was not met for the plane magnetron, where all electrons kept the same frequency of oscillation about their average (straight) trajectory.

Does this circumstance represent a serious trouble for the theory, and how does it work in practice? It depends upon the value of the resistance in the outer circuit. As already stated in the discussion of the plane magnetron, the problem discussed here corresponds to a case of infinite resistance in the outer circuit: the current is assumed to become identically naught at the end, while oscillations in the electron trajectories and in the space charges result in oscillations of the anode voltage. Infinite resistance means no damping. In an actual experiment, there will be a finite resistance in the outer circuit, hence current oscillations in the resistance accompanying the voltage oscillations, energy dissipation, and damping. After a while, the whole system will settle down on a steady state, and, if electrons do not reach a distance larger than 2.273, there is no other steady motion than the B single stream motion where electrons run on concentric circles about the cathode, with no oscillations left. According to conditions during the discharge $0 < t < T$, this final stage may be reached with or without intermediate intervals of double stream motion. If the damping due to the outer resistance works fast enough, before the double stream time intervals are reached, the final stage may be obtained directly, without any perturbation of the single stream motion. This will be the case, if the current remains very small during the discharge $0 < t < T$, since this decreases at will the rate of change of ω_2 (eq. 84).

What happens for electronic trajectories larger than 2.273 remains an open question. The double stream time intervals may again be of short duration, with trajectories finally disentangling themselves, or a final double stream motion may be obtained. The method developed here is unable to answer the question.

XIII. Small Oscillations in a Cylindrical Magnetron

The general method developed in Section X and XI can be used for an investigation of oscillations in the cylindrical one-anode magnetron. We assume a current (per unit length)

$$I = I_d(1 + \alpha \cos \omega t) \qquad \alpha \ll 1 \tag{88}$$

and introduce, according to eq. 49, the quantity

$$\epsilon = \frac{(-e)I}{8\pi\epsilon_0 m a^2 \omega_H^3} = \epsilon_d(1 + \alpha \cos \omega t) \qquad \epsilon_d = \frac{(-e)I_d}{8\pi\epsilon_0 m a^2 \omega_H^3} \ll 1 \tag{89}$$

We have then, eqs. 47 and 49,

$$\varphi = 2\omega_H \int_{t_0}^{t} \epsilon(t')dt' = \varphi_0 + \alpha\varphi_1 \tag{90}$$

with $\varphi_0 = 2\omega_H \epsilon_d(t - t_0)$

$$\varphi_1 = \epsilon_d \frac{2\omega_H}{\omega} (\sin \omega t - \sin \omega t_0)$$

and we shall obtain a solution where terms in ϵ_d and $\alpha\epsilon_d$ are kept, but higher power of ϵ_d or α are dropped. This corresponds to the assumption

$$\epsilon_d\alpha \gg \epsilon_d^2 \qquad \epsilon_d \ll \alpha \ll 1 \tag{91}$$

We must now find the quantities R_0 and R_1 defined in Section X. Starting with R_0 (eq. 52) we write

$$\begin{aligned} R_0 &= \sqrt{\varphi + \sqrt{\varphi^2 + 1}} = \rho_0 + \alpha\rho_1 \\ \rho_0 &= \sqrt{\varphi_0 + \sqrt{\varphi_0^2 + 1}} \end{aligned} \tag{92}$$

and, neglecting higher powers of α

$$\rho_1 = \rho_0'\varphi_1$$

with

$$\rho_0' = \frac{d\rho_0}{d\varphi_0} = \frac{1}{2}\left(\frac{\varphi_0 + \sqrt{\varphi_0^2 + 1}}{\varphi_0^2 + 1}\right)^{1/2} \tag{93}$$

We need computing

$$\ddot{R}_0 = \dot{\rho}_0 + \alpha\ddot{\rho}_1 = \ddot{\rho}_0 + \alpha\ddot{\rho}_0'\varphi_1 + 2\alpha\dot{\rho}_0'\dot{\varphi}_1 + \alpha\rho_0'\ddot{\varphi}_1 \tag{94}$$

According to eq. 90, $\varphi_0 \sim \varphi_1 \sim \epsilon_d$ and $\ddot{\rho}_0 \sim \ddot{\rho}_0' \sim \epsilon_d^2$ while $\dot{\rho}_0' \sim \epsilon_d$ since ρ_0 is a function of $\epsilon_d(t - t_0)$ hence the first three terms in eq. 94 are of the order of ϵ_d^2 or ϵ_d^3, and the last term is the only one to be of the first order. Using this result in eq. 53 we have

$$\ddot{R}_1 + \omega_1^2 R_1 = -\ddot{R}_0 = -\alpha\rho_0'\varphi_1 = \alpha\epsilon_d 2\omega_H\omega\rho_0' \sin \omega t \tag{95}$$

and R_1 must satisfy initial conditions corresponding to eq 48

$$t = t_0, \quad R_1 = 0, \quad R_1 = -R_0$$

when

$$\begin{aligned} \omega_1 &= 2\omega_H \\ \rho_0' &= \tfrac{1}{2} \end{aligned} \tag{96}$$

A special solution for R_1 correct to within linear terms in ϵ_d is readily obtained

$$R_1 = \alpha\epsilon_d \frac{2\omega_H\omega}{\omega_1^2 - \omega^2} \rho_0' \sin \omega t \tag{97}$$

Upon substitution of this expression in eq. 95 terms in $\epsilon_d\dot{\rho}_0'$, $\epsilon_d\dot{\omega}_1 \cdots$ appear in $\ddot{R}_1$ and are neglected as of higher orders in ϵ_d. In order to satisfy condition (96) we add to this special solution terms in

$$\begin{gathered} \frac{A}{\sqrt{\omega_1}} \cos \theta + \frac{B}{\sqrt{\omega_1}} \sin \theta \\ \theta = \int_{t_0}^{t} \omega_1(t')dt' = \theta_0 + \alpha\theta_1 \end{gathered} \tag{98}$$

that represent a solution of the homogeneous eq. 95. After determination of the coefficients A,B by means of eq. 96, we obtain

$$\begin{aligned} R = R_0 + R_1 = \rho_0 + \alpha\rho_1 + R_1 = \rho_0 - \frac{\epsilon_d}{2}\sqrt{\frac{2\omega_H}{\omega_1}} \sin \theta \\ + \alpha\epsilon_d\rho_0' \frac{2\omega_H}{\omega}\left[\frac{\omega_1^2}{\omega_1^2 - \omega^2} \sin \omega t - \sin \omega t_0\right] \\ - \alpha\frac{\epsilon_d}{2}\sqrt{\frac{2\omega_H}{\omega_1}}\left[\frac{2\omega_H\omega}{4\omega_H^2 - \omega^2} \sin \omega t_0 \cos \theta + \frac{4\omega_H^2}{4\omega_H^2 - \omega^2} \cos \omega t_0 \sin \theta\right] \end{aligned} \tag{99}$$

Next we compute eq 98 to obtain θ_0 and θ_1. We first separate in ω_1 the independant terms, and terms in α

$$\omega_1 = \omega_{01} + \frac{d\omega_{01}}{d\varphi_0} \alpha\varphi_1 \tag{100}$$

with

$$\omega_{01} = \omega_H \sqrt{2 + 2\rho_0^{-4}}$$

according to eq. 53, and we have

$$\theta_0 = \int_{t_0}^{t} \omega_{01}(t')dt'$$

$$\theta_1 = \int_{t_0}^{t} \frac{d\omega_{01}}{d\varphi_0}\varphi_1 dt' = \epsilon_d \frac{2\omega_H}{\omega} \int_{t_0}^{t} \frac{d\omega_{01}}{d\varphi_0}(\sin \omega t' - \sin \omega t_0)dt' \qquad (101)$$

Because of its oscillatory character, the first term in sin $\omega t'$ contributes only a term in ϵ_d that is negligible. The second term on the contrary yields a finite contribution since

$$\int_{t_0}^{t} \frac{d\omega_{01}}{d\varphi_0} dt' = \int_{t_0}^{t} \frac{d\omega_{01}}{d\varphi_0} \frac{1}{2\omega_H \epsilon_d} d\varphi_0$$

hence

$$\theta_1 = -\frac{\sin \omega t_0}{\omega}(\omega_{01}(t) - \omega_{01}(t_0)) = -\frac{\sin \omega t_0}{\omega}(\omega_{01}(t) - 2\omega_H) \qquad (102)$$

These expressions for θ_0 and θ_1 must be used in the general solution (99) where we note that θ_1 can be dropped in the last term that is already of order ϵ_d, but it must be kept in the second one, where

$$-\frac{\epsilon_d}{2}\sqrt{\frac{2\omega_H}{\omega_1}} \sin \theta = -\frac{\epsilon_d}{2}\sqrt{\frac{2\omega_H}{\omega_1}}(\sin \theta_0 + \alpha\theta_1 \cos \theta_0)$$

$$= -\frac{\epsilon_d}{2}\sqrt{\frac{2\omega_H}{\omega_{01}}} \sin \theta_0 + \alpha\frac{\epsilon_d}{2}\sqrt{\frac{2\omega_H}{\omega_{01}}}\,\frac{\omega_{01} - 2\omega_H}{\omega} \sin \omega t_0 \cos \theta_0 \qquad (103)$$

We have used ω_{01} and θ_0 instead of $\omega_0\theta$, since the whole term is of order ϵ_d and the correction would be in $\alpha\epsilon_d^2$, since $\alpha\varphi_1$ is in $\alpha\epsilon_d$ (eq. 90). Finally, our solution (99) reads

$$R = \rho_0 - \frac{\epsilon_d}{2}\sqrt{\frac{2\omega_H}{\omega_{01}}} \sin \theta_0 + \alpha\epsilon_d\rho_0' \frac{2\omega_H}{\omega}\left(\frac{\omega_{01}^2}{\omega_{01}^2 - \omega^2} \sin \omega t - \sin \omega t_0\right)$$
$$- \alpha\frac{\epsilon_d}{2}\sqrt{\frac{2\omega_H}{\omega_{01}}}\left[\left(\frac{2\omega_H\omega}{4\omega_H^2 - \omega^2} - \frac{\omega_{01} - 2\omega_H}{\omega}\right)\right.$$
$$\left. \sin \omega t_0 \cos \theta_0 + \frac{4\omega_H^2}{4\omega_H^2 - \omega^2} \cos \omega t_0 \sin \theta_0\right] \qquad (104)$$

This formula gives the self-consistent trajectories of electrons in a magnetron operated with a current (88) consisting of a constant direct current I_d plus small oscillations αI_d of frequency ω.

XIV. CALCULATION OF THE ANODE VOLTAGE

The anode voltage V_a is obtained by integrating the field, at a *certain instant t*, from cathode to anode, and the field itself results from eq. 47:

$$V_a = -\int_{r=a}^{r=b} E\,dr = -\frac{ma^2 2\omega_H^2}{e}\int_a^b \frac{\varphi}{r}\,dr \tag{105}$$

where $\varphi(t,t_0)$ is a function of the time of emission t_0 and of time t. We may write

$$\frac{dr}{r} = \frac{dR}{R} = \frac{\partial \log R}{\partial t_0}\,dt_0$$

calling t_0 the time of emission of electrons reaching the distance r at time t and t_0^* the time of emission of those reaching the anode $r = b$ at t:

$$\frac{(-e)V_a}{ma^2 2\omega_H^2} = v_a = \int_t^{t_0^*} \frac{\partial \log R}{\partial t_0}\,\varphi(t,t_0\,dt_0 \tag{106}$$

For oscillation purposes, all we need is the term in V_a that is linear in α, hence proportional to the oscillation current,

$$\frac{\partial v_a}{\partial \alpha} = \frac{\partial \log R}{\partial t_0^*}\,\varphi(t,t_0^*)\,\frac{\partial t_0^*}{\partial \alpha} + \int_t^{t_0^*}\left\{\frac{\partial^2 \log R}{\partial t_0\partial \alpha}\,\varphi(t,t_0) + \frac{\partial \log R}{\partial t_0}\,\frac{\partial \varphi(t,t_0)}{\partial \alpha}\right\} dt_0 \tag{107}$$

Integrating by parts we obtain

$$\frac{\partial v_a}{\partial \alpha} = \frac{\partial \log R}{\partial t_0^*}\,\varphi(t,t_0^*)\,\frac{\partial t_0^*}{\partial \alpha} + \frac{\partial \log R}{\partial \alpha}\,\varphi(t,t_0^*) - \frac{\partial \log R}{\partial \alpha}\,\varphi(t,t) + \int_t^{t_0^*}\left\{\frac{\partial \log R}{\partial t_0}\,\frac{\partial \varphi}{\partial \alpha} - \frac{\partial \log R}{\partial \alpha}\,\frac{\partial \varphi}{\partial t_0}\right\} dt_0 \tag{108}$$

Now, according to the definition of t_0^*

$$\log R\,(t,t_0^*,\alpha) = \log \frac{b}{a} \quad a \text{ cathode radius, } b \text{ anode radius} \tag{109}$$

hence
$$\frac{\partial \log R}{\partial t_0^*}\,\frac{\partial t_0^*}{\partial \alpha} + \frac{\partial \log R}{\partial \alpha} = 0$$

and the first two terms in eq. 108 cancel out. The third term is zero since $\varphi(t,t)$ and the field on the cathode are zero, and there remains only the integral of eq. 108.

$$\frac{\partial v_a}{\partial \alpha} = \int_t^{t_0^*} \frac{1}{R}\left\{\frac{\partial R}{\partial t_0}\,\frac{\partial \varphi}{\partial \alpha} - \frac{\partial R}{\partial \alpha}\,\frac{\partial \varphi}{d t_0}\right\} dt_0 \tag{110}$$

We want the derivative $\partial v_a/\partial \alpha$ for $\alpha = 0$, hence from eq. 90

$$\frac{\partial \varphi}{\partial \alpha} = \varphi_1 = \epsilon_d\,\frac{2\omega_H}{\omega}\,(\sin \omega t - \sin \omega t_0) \tag{111}$$

$$\left(\frac{\partial \varphi}{\partial t_0}\right)_{\alpha=0} = -2\omega_H\epsilon_d$$

Furthermore, we obtain from eq. 104

$$\frac{\partial R}{\partial \alpha} = \epsilon_d \rho_0{}' \frac{2\omega_H}{\omega}\left(\frac{\omega_{01}{}^2}{\omega_{01}{}^2 - \omega^2}\sin \omega t - \sin \omega t_0\right) - \frac{\epsilon_d}{2}\sqrt{\frac{2\omega_H}{\omega_{01}}}[\ \cdots\ ] \qquad (112)$$

where the bracket is the same as in eq. 104, and

$$\left(\frac{\partial R}{dt_0}\right)_{\alpha=0} = \frac{\partial \rho_0}{\partial t_0} - \frac{\epsilon_d}{2}\sqrt{\frac{2\omega_H}{\omega_{01}}}\cos \theta_0 \frac{d\theta_0}{\partial t_0}$$

but according to eqs. 90 and 101,

$$\frac{\partial \rho_0}{\partial t_0} = \frac{d\rho_0}{d\varphi_0}\frac{d\varphi_0}{dt_0} = -2\omega_H \epsilon_d \rho_0{}' \quad \text{and} \quad \frac{d\theta_0}{dt_0} = -2\omega_H$$

hence

$$\left(\frac{\partial R}{\partial t_0}\right)_{\alpha=0} = -2\omega_H \epsilon_d\left(\rho_0{}' - \frac{1}{2}\sqrt{\frac{2\omega_H}{\omega_{01}}}\cos \theta_0\right) \qquad (113)$$

Finally, we obtain from eqs. 110 to 113

$$\begin{aligned}\frac{\partial r_a}{\partial \alpha} = \frac{4\omega_H{}^2\epsilon_d{}^2}{\omega}\Bigg\{\Bigg[\int_t^{t_0^*}\frac{\rho_0{}'\omega^2 dt_0}{\rho_0(\omega_{01}{}^2 - \omega^2)} + \frac{1}{2}\int_t^{t_0^*}\sqrt{\frac{2\omega_H}{\omega_{01}}}\frac{\cos \theta_0}{\rho_0}\,dt_0\Bigg]\sin \omega t \\ - \frac{1}{2}\int_t^{t_0^*}\frac{dt_0}{\rho_0}\sqrt{\frac{2\omega_H}{\omega_{01}}}\Bigg[\left(\frac{4\omega_H{}^2}{4\omega_H{}^2 - \omega^2} - \frac{\omega_{01} - 2\omega_H}{2\omega_H}\right)\sin \omega t_0 \cos \theta_0 \\ + \frac{2\omega_H\omega}{4\omega_H{}^2 - \omega^2}\cos \omega t_0 \sin \theta_0\Bigg]\Bigg\}\end{aligned} \qquad (114)$$

where we wrote, according to eq. 99,

$$\left(\frac{1}{R}\right)_{\alpha=0} = \frac{1}{\rho_0}$$

dropping terms in $(\epsilon_d/2)\sqrt{2\alpha_H}\ \omega_{01} \sin \theta_0$ that are of high order in ϵ_d. To evaluate the integrals in eq. 114 we use the relation

$$\frac{d\theta_0}{dt_0} = -2\omega_H$$

and we write

$$\cos \theta_0 = -\frac{1}{2\omega_H}\frac{d}{dt_0}\sin \theta_0 \qquad (115)$$

$$2 \sin \omega t_0 \cos \theta_0 = \frac{1}{2\omega_H - \omega}\frac{d}{dt_0}\cos(\omega t_0 + \theta_0) - \frac{1}{2\omega_H + \omega}\frac{d}{dt_0}\cos(\omega t_0 - \theta_0)$$

$$2 \cos \omega t_0 \sin \theta_0 = \frac{1}{2\omega_H - \omega}\frac{d}{dt_0}\cos(\omega t_0 + \theta_0) + \frac{1}{2\omega_H + \omega}\frac{d}{dt_0}\cos(\omega t_0 - \theta_0)$$

Next we integrate by parts, and keep only the boundary terms, since the remaining integrals contain $d\rho_0\, dt_0$ and $d\omega_0\, dt_0$ that are proportional to ϵ_d. Hence

$$\begin{aligned}\frac{\partial v_a}{\partial \alpha} = \frac{4\omega_H^2\epsilon_d^2}{\omega}\Bigg\{\Bigg[\int_t^{t_0^*} \frac{\rho_0'\omega^2\, dt_0}{\rho_0(\omega_{01}^2 - \omega^2)} - \frac{\sin\theta_0^*}{4\omega_H\rho_0}\sqrt{\frac{2\omega_H}{\omega_{01}}}\Bigg|_t^{t_0^*}\Bigg]\sin\omega t \\ - \frac{1}{4\rho_0}\sqrt{\frac{2\omega_H}{\omega_{01}}}\left(\frac{2\omega_H}{(2\omega_H - \omega)^2} + \frac{2\omega_H - \omega_{01}}{2\omega_H(2\omega_H - \omega)}\right)\cos(\omega t_0 + \theta_0)\Bigg|_t^{t_0^*} \\ + \frac{1}{4\rho_0}\sqrt{\frac{2\omega_H}{\omega_{01}}}\left(\frac{2\omega_H}{(2\omega_H + \omega)^2} + \frac{2\omega_H - \omega_{01}}{2\omega_H(2\omega_H + \omega)}\right)\cos(\omega t_0 - \theta_0)\Bigg|_t^{t_0^*}\Bigg\}\end{aligned} \qquad (116)$$

Now we write

$$\rho_0(t,t_0^*) = \rho_0^* \qquad \omega_{01}(t,t_0^*) = \omega_{01}^* \qquad \theta_0(t,t_0^*) = \theta_0^*$$

and

$$\begin{aligned}\tau &= t - t_0^* \quad \text{transit time} \\ \omega t_0^* + \theta_0^* &= \omega t - \omega\tau + \theta_0^*\end{aligned} \qquad (117)$$

Furthermore, for $t_0 = t$ we have $\rho_0 = 1$, $\theta_0 = 0$, $\omega_{01} = 2\omega_H$ hence

$$\begin{aligned}\frac{\partial v_a}{\partial \alpha} = \frac{4\omega_H^2\epsilon_d^2}{\omega}\Bigg\{\Bigg[\qquad\Bigg]\sin\omega t + \frac{1}{2}\left(\frac{\omega_H}{(2\omega_H - \omega)^2} - \frac{\omega_H}{(2\omega_H + \omega)^2}\right)\cos\omega t \\ + \frac{1}{4\rho_0^*}\sqrt{\frac{2\omega_H}{\omega_{01}^*}}\Bigg[-\left(\frac{2\omega_H}{(2\omega_H - \omega)^2} + \frac{2\omega_H - \omega_{01}^*}{2\omega_H(2\omega_H - \omega)}\right)\cos(\omega t - \omega\tau + \theta_0^*) \\ + \left(\frac{2\omega_H}{(2\omega_H + \omega)^2} + \frac{2\omega_H - \omega_{01}^*}{2\omega_H(2\omega_H + \omega)}\right)\cos(\omega t - \omega\tau - \theta_0^*)\Bigg]\Bigg\}\end{aligned} \qquad (118)$$

where [ ] represents the same term as in eq 116.

XV RESISTANCE AND REACTANCE OF A CYLINDRICAL MAGNETRON

The internal resistance* R and reactance X of the tube are defined by

$$\frac{1}{I_d}\frac{\partial V_a}{\partial \alpha} = R\cos\omega t + X\sin\omega t \qquad (119)$$

Now, according to eqs 106 and 89,

$$\frac{1}{I_d}\frac{\partial V_a}{\partial \alpha} = \frac{ma^2 2\omega_H^2}{(-e)I_d}\frac{\partial v_a}{\partial \alpha} = \frac{1}{4\pi\epsilon_0\omega_H\epsilon_d}\frac{\partial v_a}{\partial \alpha} \qquad (120)$$

---

* The resistance R will not be confused with the ratio $R = r/a$ previously used in the discussion.

and we obtain R and X by separating terms in cos ωt or sin ωt in eq. 118

$$
\begin{aligned}
R = \frac{\omega_H \epsilon_d}{2\pi\epsilon_0\omega} \Bigg\{ & \frac{\omega_H}{(2\omega_H - \omega)^2}\left(1 - \frac{1}{\rho_0^*}\sqrt{\frac{2\omega_H}{\omega_{01}^*}}\cos(\theta_0^* - \omega\tau)\right) \\
& - \frac{1}{2\rho_0^*}\sqrt{\frac{2\omega_H}{\omega_{01}^*}}\,\frac{2\omega_H - \omega_{01}^*}{2\omega_H(2\omega_H - \omega)}\cos(\theta_0^* - \omega\tau) \\
& - \frac{\omega_H}{(2\omega_H + \omega)^2}\left(1 - \frac{1}{\rho_0^*}\sqrt{\frac{2\omega_H}{\omega_{01}^*}}\cos(\theta_0^* + \omega\tau)\right) \\
& + \frac{1}{2\rho_0^*}\sqrt{\frac{2\omega_H}{\omega_{01}^*}}\,\frac{2\omega_H - \omega_{01}^*}{2\omega_H(2\omega_H + \omega)}\cos(\theta_0^* + \omega\tau)\Bigg\}
\end{aligned}
$$

$$
\begin{aligned}
X = \frac{\omega_H \epsilon_d}{2\pi\epsilon_0\omega} \Bigg\{ & 2[\;\cdot\;\cdot\;\cdot\;] + \frac{1}{2\rho_0^*}\sqrt{\frac{2\omega_H}{\omega_{01}^*}}\left[\frac{2\omega_H}{(2\omega_H - \omega)^2} \right. \\
& \left. + \frac{2\omega_H - \omega_{01}^*}{2\omega_H(2\omega_H - \omega)}\right]\sin(\theta_0^* - \omega\tau) \\
& + \frac{1}{2\rho_0^*}\sqrt{\frac{2\omega_H}{\omega_{01}^*}}\left[\frac{2\omega_H}{(2\omega_H + \omega)^2} + \frac{2\omega_H - \omega_{01}^*}{2\omega_H(2\omega_H + \omega)}\right]\sin(\theta_0^* + \omega\tau)\Bigg\} \qquad (121)
\end{aligned}
$$

For a plane magnetron we have

$$
\omega_{01} = \omega_{01}^* = 2\omega_H \qquad \rho_0 = \rho_0^* = 1 \qquad \rho_0' = \tfrac{1}{2}
$$
$$
\theta_0^* = 2\omega_H(t - t_0^*) = 2\omega_H\tau
$$

and formula (121) for R reduces to formula (118) obtained in a preceding report. The second and fourth term cancel out, and we recognize the first positive term and the third negative one. Note, however, that

$$
\frac{1}{\rho_0^*}\sqrt{\frac{2\omega_H}{\omega_{01}^*}} = \frac{1}{\rho_0^*}\left(\frac{2}{1 + (\rho_0^*)^{-4}}\right)^{1/4} = \left(\frac{2}{(\rho_0^*)^4 + 1}\right)^{1/4} = A < 1 \qquad (122)
$$

hence the first positive term cannot be made zero for a cylindrical structure. Most favorable conditions for negative resistance are obtained, in the case $\omega < 2\omega_H$, when

$$
\omega < 2\omega_H \begin{cases} \cos(\theta_0^* - \omega\tau) = 1 \\ \cos(\theta_0^* + \omega\tau) = -1 \end{cases} \qquad (123)
$$

hence

$$
\theta_0^* - \omega\tau = 2n\pi \qquad \theta_0^* + \omega\tau = (2p + 1)\pi \quad (n,\ p \text{ are integers})
$$

These conditions are similar to those found for a plane magnetron. The minimum resistance thus obtained is

$$
\begin{aligned}
R_{\min} &= \frac{\omega_H \epsilon_d}{2\pi\epsilon_0\omega}\left\{\frac{\omega_H}{(2\omega_H - \omega)^2}(1 - A) - \frac{\omega_H}{(2\omega_H + \omega)^2}(1 + A) - A\,\frac{2\omega_H - \omega_{01}^*}{4\omega_H^2 - \omega^2}\right\} \\
&= -\frac{8\omega_H^4\epsilon_d}{\pi\epsilon_0\omega(4\omega_H^2 - \omega^2)^2}\left\{\frac{\omega}{2\omega_H} - A + A\,\frac{\omega_{01}^*}{4\omega_H}\left(1 - \frac{\omega^2}{4\omega_H^2}\right)\right\} \qquad (124)
\end{aligned}
$$

The bracket is characteristic for the resistance and for its sign. We must remember that

$$\frac{1}{2\sqrt{2}} < \frac{\omega_{01}^*}{4\omega_H} < \frac{1}{2} \tag{125}$$

For small ω the bracket is negative, but there is an interval of frequencies

$$\omega_1 \leq \omega < \omega_2$$

where the bracket is positive, and it becomes negative again for higher frequencies. The limits ω_1 and ω_2 are the roots of the equation obtained by making the bracket naught. Frequencies higher than $2\omega_H$ would require a special discussion. It must be remembered that ρ_0^* and ω_{01}^* are functions of the ratio b/a of anode-to-cathode radius only:

$$\rho_0^* = \frac{b}{a} \qquad \omega_{01}^* = \omega_H \sqrt{2 + 2(a/b)^4} \tag{126}$$

It would be interesting to discuss the type of electron trajectory corresponding to conditions (123) and to see whether it reproduces the very special features observed for the plane magnetron. Another problem is the behavior of magnetron with large ratios of anode to cathode radius, when $\rho_0^* > 2.273$, and the static characteristic of Fig. 9 is obtained. The question would be to investigate how negative resistance reappears for high frequency, while it does not exist for very low frequencies, at least as far as the static characteristic is concerned.

REFERENCES

1. Brillouin, L. Magnetron I. *Phys. Rev.*, **60,** 385 (1941).
2. Brillouin, L. Practical results from theoretical studies. *Proc. Inst. Radio Engrs.*, **32,** 216 (1944).
3. Allis, W. P. Theory of the Magnetron Oscillator. M.I.T. Radiation Laboratory, Report V-9S, October, 1941.
4. Allis, W. P. Outline for a Theory of Space Charge. M.I.T. Radiation Laboratory, Report 43-3, July, 1942.
5. Slater, J. C. Theory of the Magnetron Oscillator. M.I.T. Radiation Laboratory, Report V-5S, August, 1941.
6. Slater, J. C. Theory of Magnetron Operation (theory of scaling). M.I.T. Radiation Laboratory, Report 43-28, March, 1943.
7. Brillouin, L. Cylindrical Magnetron. A.M.P. Report 129. 2. R- July, 1944.
8. Hartree, D. R. British Report C. V. D. Mag. 1, 1941; 3, 1941; 6, 1941; 12, 1942; 23, 1942; 30, 1943.
9. Stoner, E. C. British Report C. V. D. Mag. 8, 1941; 16, 1942; 17, 1942; 25, 1942.
10. Bloch, F. Two notes from the Radio Research Lab., Harvard University, Feb. 7, 1945 and June 11, 1945.
11. Brillouin, L. A theorem of Larmor. *Phys. Rev.*, **67,** 260 (1945).

Tube Miniaturization

JOHN E. WHITE*

National Bureau of Standards, Washington, D.C.

CONTENTS

Page
I. Introduction 183
II. Limitations in Miniaturization of Tubes 184
  1. Expected Operational Shortcomings 184
    a. Short Life 184
    b. Microphony 185
    c. Unreliability 185
  2. Physical Sources of Limitation 186
    a. Current Density 186
    b. Cleanup 186
    c. Anode Dissipation 187
    d. Envelope Dissipation 187
    e. Grid Emission 187
    f. Mechanics of Fabrication 188
  3. Economics 189
III. Noteworthy Features of Subminiatures 189
  1. Special Grids 190
  2. Cathode Techniques 190
  3. Anode Materials 191
  4. Envelope Facts 191
  5. Solutions of the Sealing Dilemma 192
IV. Summary State of the Art 192
  References 194

I. INTRODUCTION

As indicated in the accompanying article, electronic components of minimum dimensions are attaining wide use and great importance. It may be stated confidently that the rapid acceptance of tiny tubes in recent years is but a token of enormously increased acceptance in the near future, as excellent performance is recognized and still better performance achieved.

The "miniature" tube first appeared in 1939; by the end of World War II, over 50 million of these tubes had been used.[1] These "miniatures" are reduced to 1.9 cm maximum diameter from 3.3 cm for equivalent GT's;** and to 5.4 cm maximum length from 8.4 cm for GT's. The trend now is to go still farther, to the so-called subminiature sizes, whose largest version uses the T-3* bulb.

---

* Present address: General Electric Co., Schenectady, New York.

** GT, for "Glass Tube," is a designation adopted for the standard radio receiving tube with glass envelope to distinguish from the MT, or "Metal-Envelope Tube."

It is the purpose of this paper to give an indication of the functional variety now available in highly miniaturized tubes, to point out operational and physical difficulties and limitations encountered as size is reduced, and to describe technological methods being employed to overcome the difficulties.

Since there are now over one hundred types of subminiature tubes commercially available, those mentioned below are merely indicative of the scope covered by this new tool.

No part of the paper is devoted to cataloguing the "advantages" of subminiatures, although some space is given to their limitations. This is because, in general, the user wants just one advantage, which is small size; he is willing to accept certain disadvantages to gain this end.

Actually, however, very small tubes do have certain advantages intrinsic in their size. One of these advantages is a generally higher frequency limit of operation obtainable with scaled-down dimensions, due to transit time and inductive effects. Another not inconsiderable advantage is the increased ruggedness provided by small structures. It is easier to make subminiature tubes to withstand high shock and vibration forces without damage than to make larger tubes for the same service.

II LIMITATIONS IN MINIATURIZATION OF TUBES

Before giving way to enthusiastic plans for reducing all types of tubes to minimum dimensions, stock should be taken of the drawbacks to be encountered and the extent to which these are fundamental or can easily be overcome. From the standpoint of the user, such drawbacks are essentially operational or economic, whereas the designer and the student of physical processes are more concerned with reasons why there are limitations. The three viewpoints are considered separately in subheadings of this chapter.

1. *Expected Operational Shortcomings*

a. Short Life. Relatively high ratings per unit volume are the rule in subminiatures and early lots were very short lived. Improved processing, cathode studies, and quality control methods have now increased subminiature life to where it is competitive with that of other tubes for the same class of service. One manufacturer now claims a life expectancy of 2500 hours for most of his subminiatures and 5000 hours for premium tubes.

---

* The T size indicates *nominal* bulb outside diameter, in eighths of an inch. Thus, a T-1 has a nominal OD of $^1/_8$ inch, or nearly $^1/_3$ cm; a T-3 bulb is nominally just under 1 cm in outside diameter, etc.

A life expectancy of t hours does not, of course, imply that all tubes with that expectancy will hold their characteristics that long. A standard finding acceptance in the United States calls for any "lot" (or batch) of tubes of a given type to have an average life not less than 80% of that specified for the type. End of life for any tube occurs when its characteristics deviate the maximum permissible amount from their specified values.

Statistical studies of tube life and causes of failure are still in their very early stages, even for conventional tubes, so that an accurate picture of life in actual service is not possible at present. This unsatisfactory state of affairs has arisen because the cost of shipment, study of service conditions, and fault isolation in tubes which have failed while performing their duties would in general exceed the cost of original manufacture.

b. Microphony. Subminiature tubes are frequently forced to operate under conditions of shock and vibration; the electrical output caused by such mechanical stress on a tube is called microphony. Although, as mentioned earlier, small tubes will in general withstand higher acceleration than the large ones without permanent damage, smallness does not give any advantage microphonically. This is because the mechanical displacement of internal parts which are already crowded together produces a larger electrical effect than the same displacement when spacings are larger.

Microphony is being reduced by making parts stiffer and fits snugger and by locating the parts more accurately with respect to each other. A number of very small tubes are now quite adequate microphonically for their intended applications.

c. Unreliability. The recent historical period, which has seen the birth and early growth of the subminiature tube, has also witnessed a more and more insistent demand for tubes of unquestioned reliability. The requirements of electronic computers, aircraft control systems, and instrumentation, for example, are very severe and cannot compromise with tubes prone to frequent failures. It has been necessary in the past for users with highest quality requirements to avoid, sometimes with reluctance, the use of subminiature tubes. One major factor here was questionable life expectancy, discussed in the preceding section. Uniformly high quality assembly is difficult to achieve in very small sizes where parts and aggregates must be inspected under magnification; loose welds and the like are to be anticipated in relatively high proportion. Absolute dimensional tolerances must be tightened considerably in small tubes to achieve even approximately the same degree of reproducibility of electrical characteristics as commonly obtained in the GT sizes.

Despite all this, subminiatures are now being offered for instrument service.[2] More accurate jigging; more thorough inspection; and designs less susceptible to assembly errors will undoubtedly make the subminiature tube the cornerstone of electronic instrumentation someday. At present, however, the quality tube user is safer with larger sizes, unless he has an unusually good record of the performance of the subminiature type he contemplates using.

2. *Physical Sources of Limitation*

Progress in achieving acceptable performance with minimum size is the result of successfully coping with a number of physically limiting factors Those of especial significance will be discussed in the paragraphs below as problems, whose solutions will appear, in the main, in Chapter III.

a. Current Density. Oxide-coated cathodes are not usually called on to deliver more than 0.2 amp per square centimeter average in tubes expected to have a reasonably long life. Subminiatures are, of course, subject to this limitation, so that current obtainable depends in part on the cathode area which can be provided in the space available. This area is quite small per unit length when the cathode is a coated wire filament; this type of cathode is preferred in hearing-aid tubes because of low drain required from the batteries used as power supplies.

Current density also restricts the size to which openings may be reduced in gas tubes, such as thyratrons, if current requirements are fixed. Several amperes average per square centimeter is a reasonable design target in ordinary gas tubes, but increased pressures permit relaxation of this figure in subminiatures, where the increased pressures are permissible because of reduced electrode spacings. Existing subminiature gas tubes have not gone as far as possible in this direction.

Figure 1 shows the arrangement of parts in a subminiature thyratron, indicating the degree of constriction imposed on the conducting path by the control grid. The tube of Fig. 1 is rated at 100 ma peak.

b. Cleanup. Gas pressure change during operation is a perennial difficulty in permanent-gas-filled tubes. This problem of cleanup would seem more serious in subminiatures than in large tubes because ratings tend to vary as the square of the linear dimensions while volume varies as the cube. The quantity of gas available varies as the volume, for like pressures. The saving fact here is that higher gas pressures may be used in small-dimension tubes, so that cleanup is really no more difficult to combat in small tubes than in large ones.

c. Anode Dissipation. As in larger tubes, most of the power loss must be radiated from the anode. This electrode is generally operated well below incandescence, largely to avoid the necessity of an economically unfeasible degree of outgassing during manufacture, but also for secondary reasons such as avoidance of evaporation of the anode material, thermal buckling, and excessive heat conduction to seals. The anode radiating area cannot be reduced below that necessary to dissipate rated losses at a permissible temperature

d. Envelope Dissipation. The envelope of the tube is cooled by both radiation and convection and obviously must not operate at a temperature too close to the exhaust bake-out temperature, which is generally about 350°C for subminiatures with soft glass envelopes. In some applications, high envelope temperatures also result in detrimental glass electrolysis at the leads About the maximum temperature presently permitted for subminiature envelopes is 250°C.

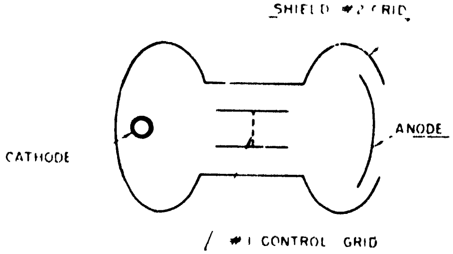


FIG. 1. Electrode arrangement in a subminiature thyratron with 1-cm outside envelope diameter. (*Courtesy Sylvania Electric Products, Inc.*)

e. Grid Emission Close spacings and high operating temperatures make grid emission a serious problem in some of the very small tubes. Emission coating evaporates from the cathode in all tubes with activation and operating temperatures generally used, and small tubes generally do not have the space required to shield the grid from this material. A grid coated with emission compounds will release enough thermionic electrons to trouble many circuits at quite moderate temperatures. Exposure of the grid to radiation from the cathode increases the thermal problem, while cooling is not effective with the limited-area side rods used to support the grid. Evaporation from the cathode is accentuated in applications where the cathode heater voltage varies over a considerable range. It is clear that reduced ratings, low cathode temperatures, and narrow tolerances on cathode heater voltages will ease the grid emission problem; but it is also clear that design considerations are in order. These will be taken up in Chapter III.

f. Mechanics of Fabrication. Most subminiature tubes are designed and manufactured like their larger counterparts. Cathode coating, grid winding, bulb sealing, and exhaust processing are generally carried out on automatic machinery; but assembly or mounting operations are done by hand, frequently without benefit of magnification. It may be that the ultimate has not yet been reached in this type of fabrication. But the prediction that production costs will become prohibitive in this "watchmaker's art" before sizes have been farther reduced seems reasonable.

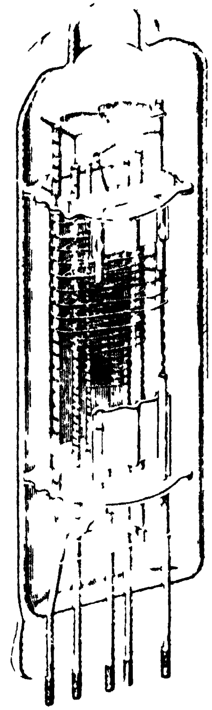


FIG. 2. Cutaway view of a subminiature pentode. The bulb cross-section dimensions are approximately $^2/_3$ × 1 cm (T2 × 3), outside. (*Courtesy Raytheon Manufacturing Co.*)

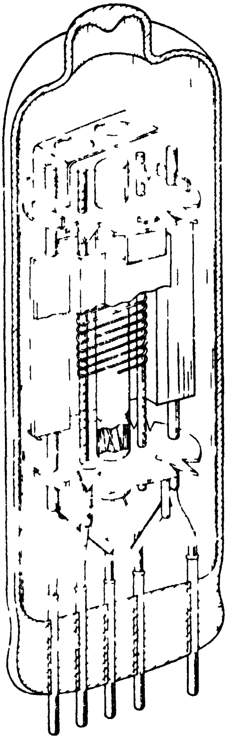


FIG. 3. Subminiature UHF triode. (*Courtesy Raytheon Manufacturing Co.*)

A remarkable fact is that low-cost, high-production fabricated and stamped parts are held in close-spaced, mutually insulated positions with an accuracy of location closer than usual machining tolerances. Some fixtures are used, but not on all tubes. The basic jigs are really the mica wafers built into the tube at each end of its structure, with accurately placed holes to locate the parts. A serious size and complexity limitation is, in fact, the minimum spacing attainable between mica holes without breakthrough, this being about 0.03 cm for usual mica thickness.

Figure 2 is worth a thousand words in showing the complexity contrived into a typical subminiature pentode. Coaxial elements include a filament, three grids, and a plate. The filament is held taut by a cantilever spring shown near the top and is centered by an arrow cut through the mica. Dark circles in the "flag" at the top are getter compound.

Figure 3 is another tube of the same bulb size as shown in Fig. 2; it illustrates the use of an indirectly heated cathode in this size tube.

A major fabrication problem in the manufacture of small tubes is the sealing of the mount into the bulb. This is more difficult than in large tubes because it is difficult in the limited space available to avoid oxidizing the parts and poisoning the cathode. American practice calls for automatic sealing in a protective gas atmosphere, with preheat, sealing and anneal fires delicately adjusted to give just the heat needed to get vacuum tightness and prevent excessive glass strains with resulting cracks, without overheating the parts. The tubes shown in Figs. 2 and 3 are sealed in by supporting the mount independent of the glass and heating the glass to softness, then quickly pressing the glass around the leads to form a seal. Another American method is to mount the parts on a glass button, prepressed or sintered around the leads. The tube envelope and exhaust tubulation assembly is then dropped over the mount and sealed to the button. Other methods of sealing will be discussed in Chapter III.

3. Economics

The cost aspect should be mentioned briefly, for the benefit of those who plan applications. Since the output capabilities of a tube go down with size but its complexity does not, it seems rather fundamental that paralleled subminiatures as now designed cannot compete economically with equivalent rated single tubes of larger size.

III. NOTEWORTHY FEATURES OF SUBMINIATURES

Design, materials, and methods now used in subminiatures are mostly adaptations or extensions of the same in larger tubes. This situation may be expected to change as the small types develop out of infancy. Peculiarly adaptable features are already finding broad application in size reduction and new techniques are being explored. Some of the more interesting or newer features will be discussed in this section.

1. *Special Grids*

Wound grids, traditionally of nickel, are generally made of the more refractory materials where small wire diameters are required. Molybdenum, tungsten, and various alloys containing these metals are predominant. In a process finding some application,[3] the desired grid mesh is produced directly by electroplating, instead of using wires.

Under certain conditions, coatings on the grid of gold,[4] silver,[5] or other materials[6] may be expected to reduce grid emission, which has already been mentioned as a problem.

Planar microwave triodes have developed some miniaturization techniques of their own to achieve small electrode spacings and accordingly small electron transit times. Most of these tubes use grids of fine wire wrapped under tension on and brazed to a supporting ring. Ring spacers around the edges of the elements are used to set the required spacings. The most elegant tube of this type yet produced is the Bell Labs 1553,[4] designed to generate up to 4000 megacycles per second. A comparison of the spacings in this tube with those of a planar tube released for production several years ago is shown, to scale, in Fig. 4. The cathode-grid spacing in the 1553 is only 1.5×10^{-3} cm; the grid wires are 7.5×10^{-4} cm in diameter, wound at 395 turns per centimeter.

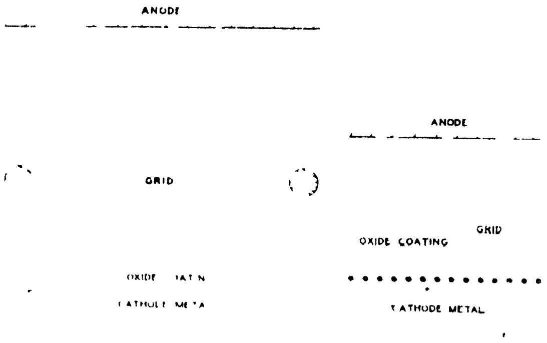


FIG. 4 Comparative size reduction in planar microwave triodes. Lighthouse tube on the left Grid-cathode spacing in the 1553 is 1.5×10^{-3} cm (*Courtesy Bell Telephone Laboratories.*)

2. *Cathode Techniques*

The cathode coating on the 1553, just mentioned, is applied by an automatic spray machine, which applies oxide to the metal base in a layer $1.3 \times 10^{-3} \pm 5 \times 10^{-5}$ cm thick.[4]

Accurate control of the cathode coating is sometimes obtained by cataphoretic[7] application.

The base metal of sleeve cathodes is generally nickel, as in larger tubes. In filamentary cathodes, however, nickel is not strong enough in the small sizes required for some applications; especially in hearing-aid tubes where filament drain must be kept low to avoid depleting the small battery used as a source of power. Alloys stronger than nickel at operating temperatures, such as Ni-Cr alloys or refractories such as tungsten, are used as filaments in many tubes, with the conventional alkali-earth oxide coatings. Tungsten filaments only 8×10^{-4} cm in diameter and requiring a heating current of only 12.5 ma are in use.[8]

3. *Anode Materials*

Carbonized nickel, with its relatively high thermal emissivity, is a common anode material in subminiatures. Operation of the anode at temperatures above the start of incandescence would allow dimensions of the plate to be reduced, with further possible advantages from using plate materials with natural gettering properties such as tantalum or zirconium. Overall size reduction may be accomplished in certain tubes by such increased specific loading of the plate; but in general, plate size reduction will not help until the envelope size may also be reduced. This is also possible, as shown in the next section.

One plate material developed during World War II in Germany is now being applied to some extent in small tubes. This is aluminum-coated iron sheet which, heated in vacuum, takes on a highly emissive black surface with gettering properties.[9]

4. *Envelope Facts*

As mentioned before, most subminiatures are made with soft-glass envelopes and dumet seals. It would seem possible to reduce the size of such tubes appreciably by using harder glass and higher bake-out temperatures. This would make a more expensive tube, a factor tolerable in some applications. A more fundamental limitation preventing wide adoption of this expedient is that with the small glass thicknesses now existing between leads, further size reduction would in most cases produce excessive electrolysis of the glass or radio-frequency losses with hard glasses just as with soft glasses. Currently used bake-out temperatures could be raised somewhat even for soft glasses, but the limit for hard kovar-sealing glass is only about 100 centigrade degrees higher than for soft glass.

A more hopeful method of achieving high envelope dissipation in small sizes is offered in the use of ceramic envelopes. Ceramics with less electrolysis and radio-frequency losses than glass are readily available. Metals are now being sealed vacuum tight directly to ceramics without the use of glass by several techniques[10] developed during and since World War II.

5. Solutions of the Sealing Dilemma

Inasmuch as damage to the cathode and other parts during sealing becomes a more serious problem as size is reduced, it is worth while to mention additional techniques for combating this difficulty.

One such method is the use of a low-melting "solder glass" which matches the envelope and button base material in thermal expansion.[8] This solder glass is sealed at 450°C, as opposed to 800 or 900°C required for the usual soft glass.

Another method finding some application where the final seal is a metal-to-metal joint is the gold diffusion technique.[3] A gold wire is clamped between two copper flanges. The joint is brought to 400°C during exhaust and a vacuum-tight diffusion seal, equivalent to a hard-soldered joint in strength and permissible operating temperatures, is produced.

A glass-to-glass seal may also be produced at normal exhaust temperatures if the surfaces are first optically polished.[11]

IV. SUMMARY STATE OF THE ART

Over a hundred subminiature tube types are already on the market. The varieties which will be produced can be expected to keep up fairly well with expanding requirements since, as demonstrated above, means are at hand for overcoming most of the factors now limiting advancement in this field.

A few more existing types should be mentioned to help orient the reader on the state of the art at the present time.

The 5647 (Sylvania) is a simple diode, but is worth mentioning because it has a 5000-hour life expectancy, with a T-1 ($^1/_3$ cm) bulb.

A construction now used by RCA[9] for several types is shown in cross section in Fig. 5. This triode oscillator will deliver over 50 mw at 3000 megacycles per second; the diameter of the body is just that of an ordinary lead pencil.

Voltage regulator tubes are available in subminiature sizes for various ratings. The 5841 (Victoreen) is noteworthy in regulating the relatively high tension of 900 volts, using a corona discharge.

Sylvania's 5642 is remarkable as a rectifier in that it incorporates a 10-kv rating in a T-3 bulb.

Photo tubes are available in subminiature sizes. An example is the RCA 1P42.

A Geiger tube, with No. 2 sewing needle for size comparison,[12] is shown in Fig. 6.

Continuing into somewhat unusual types, we find the RCA 5734 transducer. This is a subminiature triode with one electrode movable through a metal diaphragm. It may be used to convert mechanical vibrations up to 12,000 cycles per second into electric current variations.

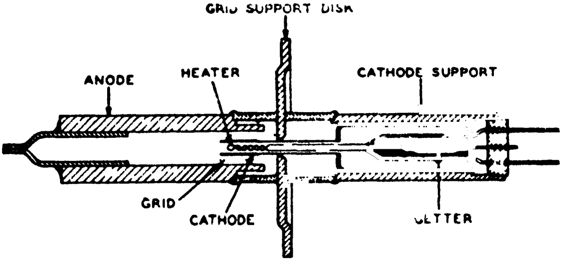


FIG. 5 Cross section of an RCA "pencil" triode (*Courtesy Radio Corporation of America.*

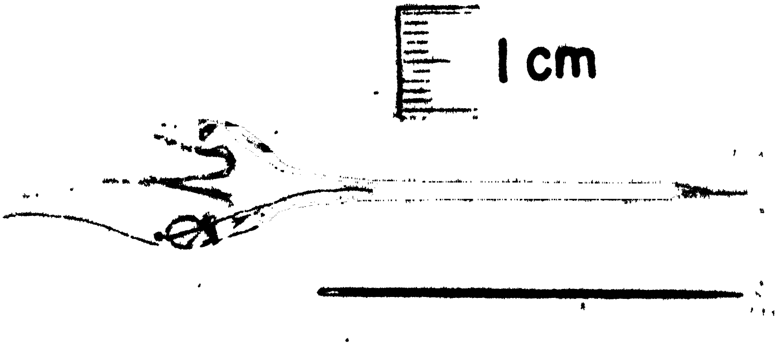


FIG 6. Miniaturized Geiger tube. (*National Bureau of Standards.*)

Finally we have the solid state devices which are electronic without being vacuum tubes and which certainly are highly miniaturized. The diode rectifier has been known for a long time; mixers, up to tetrodes, are a logical extension. The Bell Labs transistor[13] is a triode amplifier which has done much to assure the important role of solid state devices in the tube field, especially where miniaturization is of importance.

Although all the references to foreign publications cited above deal with techniques, it should be mentioned specifically that Europe is now in production on subminiature tubes.[14]

Since selected types may already be obtained with life and reliability equivalent to that of larger sizes, the major obstacles to broad acceptance of subminiatures have been removed.

Indeed, one prominent manufacturer now advertises a line of subminiatures said to outperform in every way the corresponding miniatures.

If the author may venture a prediction, however, the next major step in size reduction and reliability improvement will be based on new design concepts, permitting a radical departure from the complex assembly procedures typified in Fig. 2.

ACKNOWLEDGMENTS

Figure 1 was furnished by Sylvania Electric Products, Inc.; Figs. 2 and 3 by the Raytheon Manufacturing Co.; and Fig. 6 by Dr. Curtiss of the National Bureau of Standards. Permission to reproduce Fig. 4 was given by Bell Laboratories, and Fig. 5 by RCA. To these and to the Victoreen Instrument Company, I am indebted for supplying material useful in preparing the article. W. J. Weber of the Bureau of Standards assisted in surveying available types.

REFERENCES

1. Green, N. H. *RCA Rev.*, **VIII**, 331–340 (1947).
2. Victoreen, J. A. *Proc. Inst. Radio Engrs.*, **37**, 132–141 (1949).
3. Davies, J. W., Gardiner, H. W. B., and Gomm, W. H. *Proc. Instn. Mech. Engrs. (London)*, **158**, 352–368 (including discussion) (1948).
4. Morton, J. A. *Bell Labs Record*, **XXVII**, 166–170 (1949).
5. Rose, G. M., Power, D. W., and Harris, W. A. *RCA Rev.*, **X**, 321–338 (1949).
6. Sorg, H. E., and Becker, G. A. *Electronics*, **18**, 104–109 (1945).
7. Biguenet, C., and Mano, C. *Le Vide*, **2**, 291–304 (1947).
8. Alma, G., and Prakke, F. *Philips Tech. Rev.*, **10**, 289–295 (1946).
9. Harrison, J. S., Britten, L. F., Darlaston, A. J. H., Martin, S. L., and Wolfson, H. *BIOS Rep.* (H. M. Station. Off.) No. 1831, 1-29, 1948.
10. Bondley, R. J. *Electronics*, **20**, 97–99 (1947).
11. Danzin A., and Despois, E. *Ann. Radioélectricité*, **III**, 281–289 (1948).
12. Curtiss, L. F. *J. Research Natl. Bur. Standards*, **30**, 157 (1943).
13. Bardeen, J., and Brattain, W. H. *Phys. Rev.*, **74**, 230-231 (1948).
14. British Subminiature Valves. *Wireless World* **54**, 80–81 (1948).

Subminiaturization Techniques

GUSTAVE SHAPIRO

National Bureau of Standards, Washington, D.C.

CONTENTS

Page
I. Introduction 195
  1. Need for Subminiaturization 195
  2. Subminiaturization Defined 196
II. Design Philosophy 196
III. Thermal Considerations 197
IV. Assembly Techniques 198
  1. Printed Circuits 198
  2. Piece Components 199
  3. Plastic Embedment 200
V. Subminiature Assemblies 201
  1. Handie Talkie 201
  2. Broadcast Receivers 201
  3. Hearing Aid Amplifier 201
  4. Counter 202
  5. Proximity Fuze 204
  6. Potted Amplifier 204
  7. I-F Amplifiers 204
    a. NBS Model 2 205
    b. NBS Model PC 1 206
    c. NBS Model 5 207
VI. Components and Materials 208
  1. Insulating Materials 208
  2. Capacitors 209
  3. Conductive Paint 213
  4. Resistors 213
  5. R-F Inductors 214
  6. Transformers 214
  7. Batteries 217
VII. Outstanding Problems 217
VIII. Conclusions 218
References 218

I. INTRODUCTION

1. *Need for Subminiaturization*

The quantity of electronic equipment being used by industry and by the military has increased greatly in recent years. Often the size of a complex electronic device is so large that space cannot be provided to house it. Efforts are now being made to develop smaller, more reliable components as well as subminiaturization techniques for assembling these small components into compact, rugged, easily manufactured electronic assemblies. Two hundred seventy thousand hearing aids alone, valued at $16,868,000, were shipped by manufacturers during the year 1947, according to the Bureau of the Census, U.S. Department of Commerce.[1] Production figures of subminiature electronic units for military use are known to be very substantial but specific values are not available.

When subminiature vacuum tubes became available the need for new and radical equipment assembly techniques became apparent. For a number of years various development groups have been seeking subminiature assembly techniques with good mass production possibilities. This article will attempt to describe some of the more promising methods that have been developed and the technical advances in components and materials that make possible these subminiature assemblies.

Because the preponderance of activity has been in the United States and because so little published information on European subminiaturization development is available, the techniques to be described are derived mainly from American sources.

2. *Subminiaturization Defined*

To differentiate between the usual methods of miniature construction and the newer techniques to be described below the latter will be referred to as "subminiaturization." Subminiaturization may be defined as *Techniques that make possible an electronic assembly whose volume is compacted to a dimensional limit primarily imposed by the smallest available electron tube.* It should be observed that the term "assembly" rather than equipment is used in this definition because it is recognized that equipment may be made up of conventional as well as subminiature assemblies. By comparing the volume with that of "the smallest available electron tube," the definition becomes less restrictive and should resist obsolescence as advances in the subminiaturization and vacuum tube arts make smaller assemblies possible.

II. DESIGN PHILOSOPHY

In the past subminiaturization has been considered only a secondary objective in the design of an electronic equipment. Because little time and little effort was expended the degree of subminiaturization was never marked. This attitude rather than its inherent difficulty has retarded the development of the process.

The approach in electronic subminiaturization must be quite different from that in ordinary electronic development work. The major difference is brought about by the scarcity of suitable, available component parts. These subminiature components must operate to higher ambient temperatures than conventional size components yet still retain adequate electrical efficiency. If the subminiaturized device is to be more than just a laboratory curiosity, the components must be designed so that industry may easily tool for their production. If possible, sufficient commercial interest must be aroused so that future procurement of these new parts is no problem.

Having first assembled all available applicable subminiature components, the engineer next surveys the electronic device to be subminiaturized for the purpose of redesigning the circuits to use the fewest possible parts. Where an original circuit uses a component that does not lend itself to subminiaturization, that particular type of component[2] should be eliminated where practicable, by suitably modifying the circuit. After the assembly has been simplified to use the fewest number of component parts, and the smallest parts with the most desirable shape factor have been collected, these parts should be assembled into the most compact arrangement possible. The more flexible the shape factor of the parts, the more efficiently may subminiaturization be accomplished. Frequently, an otherwise satisfactory component must be redesigned into a new shape for a particular assembly.

Effectiveness of subminiaturization is often coupled with unitization for ease of maintenance. This facilitates replacement of a defective unit or plug-in assembly. To guard against the effects of humidity, it is generally desirable to seal hermetically the individual units comprising an assembly or to seal the assembly as a whole. Embedding an electronic assembly in a casting resin may serve as an alternate to sealing a unit hermetically.[3]

III. THERMAL CONSIDERATIONS

Subminiature assemblies can be divided into two general categories, those that operate at moderate temperatures and those that operate at higher temperatures, up to 200°C. Most military component specifications call for performance up to 175° 200°C. These are by no means ultimate desired temperatures, but represent merely a limit where solder assembly techniques may still conveniently be used.

Battery-powered equipment comes within the first category while most other equipment falls into the second. Because of the more difficult technical problems encountered in designing subminiaturized equipment that must withstand high temperatures, practical subminiaturized battery-powered equipment was produced commercially long before other types.

Any arrangement for cooling an electronic assembly should be an integral part of the electronic installation and must be considered in the overall equipment design. One of the simplest systems that could be used is forced-air circulation in conjunction with radiating fins on those enclosed assemblies that generate considerable heat. If the enclosed assemblies are filled with a fluid of high thermal conductivity, the effect will be to reduce hot spots on components, especially resistors and tube envelopes. This fluid should have a low dielectric constant and low losses. Its vapor pressure should be low so that high internal pressures will not build up at high ambient temperatures, and it should be sufficiently inert so that it will not attack the components. If the fluid is pumped through the assembly, the need for a low vapor pressure becomes less important; however, the mechanical difficulties of designing such a unit so it may plug into a circulatory system are considerable. A less efficient method, but one with fewer mechanical problems, would be to employ a hollow metal block through which a coolant is pumped. Plug-in assemblies would be thermally bonded to this cooling block.

IV. ASSEMBLY TECHNIQUES

1. *Printed Circuits*

In the early development stages of the printed circuit art, some extravagant claims were made. One of these maintained that subminiaturization could only be accomplished through the medium of printed circuits. The wide variety of subminiaturization techniques now being used disproves this statement. Printed circuits are primarily production techniques and as such should be used in subminiature assemblies where they are capable of improving fabrication problems. Generally, the best examples of subminiature assemblies are found to be those that employ a combination of printed circuits and piece components. The ratio of printed circuits to piece components may be expected to depend largely upon the degree of advancement of the printed circuit art.

Some equipment manufacturers prefer to restrict their operations to those of assembly and have no desire to become processors of components. Those manufacturers can still function in the manner they choose while taking advantage of the new printed circuit techniques by utilizing printed wiring in conjunction with piece resistors, capacitors, etc.

Although the stencil screen process[4] has had widest application to subminiaturization, other printed circuit processes are currently meeting with some success. Chief among these is a technique in which copper foil is bonded to a sheet of silicone or Teflon-impregnated fiber glass. (Teflon is the trade name for polytetrafluoroethylene.) With the exception of the desired circuit pattern, all of the metal surface is removed by means of either the photographic-acid resist or photographic-sandblast process. Some commercial companies are exploiting the possibilities of this process by incorporating small bonded copper foil r-f assemblies in their television receivers.

2. *Piece Components*

One parts manufacturer has developed a piece component counterpart of the printed electronic subassembly. These small molded networks (Fig. 1) bridge the gap between conventional assembly methods using hookup wire and those using printed circuits. R-C circuits, L-C circuits, and vacuum tubes have all been encased successfully. Prior to molding, the various piece components may be interconnected, using a spot welding technique so that the only wires coming out of the assembly are those required to connect the molded assembly into the external circuit. Cylindrical structures that fit over the associated vacuum tube can also be molded. Shielding between various network elements may be introduced when desired.

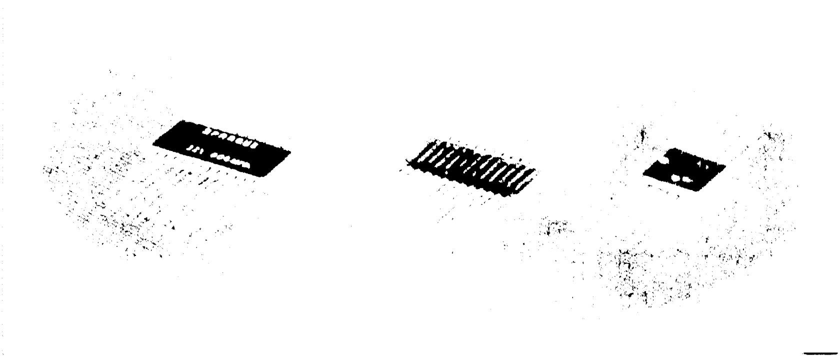


FIG 1. Piece component assemblies molded in a thermal setting resin. The central assembly has been cut away along the center line (by courtesy of Sprague Electric Co)

The catacomb assembly method, a subminiaturization technique, has met with some success. The chassis for this assembly method is a thick block of steatite honeycombed with parallel holes into which may be slipped the piece components of which the assembly is composed. The surface of the steatite block, to which the holes are normal, are silvered with the necessary interconnecting wiring. The piece components, tubes, etc., are merely dropped into their respective wells and soldered to the wiring pattern at the mouth of the well. This procedure makes extreme finger dexterity on the part of the assembly personnel unnecessary, even for the handling of tiny piece components. A variation of this technique is described in connection with the National Bureau of Standards i-f amplifier Models 2[5] and 5 to be discussed later. In these instances inductor forms with wide end flanges were used instead of solid steatite blocks.

3. Plastic Embedment

Molding and potting in an insulating medium are effective methods for protecting electronic assemblies from moisture, dust, etc. It is usual to provide some mechanical support to hold the components in position during the potting or the molding process. If the assembly is a simple one, the mechanical supports may be external to the sealed assembly, and may be used over again. In more complex assemblies, the supports are usually molded or potted with the electronic components. A commonly employed supporting technique is to insert the components into holes in a molded block of the same material used for the overall jacket. Assemblies are sometimes cast in self molds which then become the outer covering for the final assembly. If the assembly is to be unshielded, the mold may be a shell of the same material as the potting resin or it may be a material to which the potting resin will adhere. If the assembly is to be shielded, the components may be placed in a shield can and the plotting material poured in, or the plastic itself may be plated.

Most potting compounds shrink somewhat, hence it is usually necessary to place the tube envelope in a resilient cushion to take up the shrinkage strains. These strains may occur as long as a year after the assembly has been potted and may cause the tubes to crack. Since the simple expedient of slipping a piece of vinyl tubing over the vacuum tube often proves inadequate, satisfactory tube protective jackets for rigid resins may become quite elaborate and bulky, making subminiaturizing an assembly more difficult.

There are a wide variety of potting compounds with varying characteristics. For applications where low loss is paramount, one of the more desirable materials is the National Bureau of Standards casting resin.[6] This compound has electrical properties similar to those of polystyrene and has a low shrinkage factor.

Potting with thermoplastic materials is generally limited to those assemblies that do not generate much heat. This would include battery-powered equipment and those a-c powered equipments that may be broken up into units of a few tubes with total power dissipations low enough to prevent softening of the casting resin.

V. Subminiature Assemblies

1. Handie Talkie

An excellent example of a battery-powered subminiature assembly is a Handie-Talkie unit manufactured for use by firemen, police, etc. This equipment is a two-way crystal-controlled F-M communication set. It is constructed entirely of plug-in subminiature assemblies. This makes repair rapid since it is simply a matter of removing the inoperative stage and plugging in a replacement. The components are conventional miniature types, interconnected in the customary manner with hook-up wire. The resistors are tiny radial-lead units This device is a good example of a controlled subassembly shape factor which yields a complete unit with a high degree of space utilization.

2. Broadcast Receivers

At least two manufacturers have designed pocket personal superheterodyne broadcast receivers. In one of these an attempt has been made to keep the cost down by incorporating a few commercial resistor-capacitor printed subassemblies; the other uses miniature components throughout. One employs a loop antenna and a loudspeaker reproducer. The other uses a hearing aid reproducer in which the reproducer cord performs the additional function of an antenna.

3. Hearing Aid Amplifier

Portable hearing-aid devices reached a high degree of technical excellence with the introduction of a compact, completely printed three-stage audio amplifier (Fig. 2). A ceramic plate, approximately 1⅛ in. by 2¼ in., mounts 7 printed resistors, 6 capacitors, 3 vacuum tubes, and the necessary printed interconnections. The resistors are phenolic bonded, ⅕-watt printed elements, the capacitors are thin silvered disks of barium titanate, and the connective paths are fused metallic silver, which adhere to the ceramic base plate with a tensile strength of 3000 lb. per square inch. This unit is typical of the class of printed circuits that may be produced on a monoplanar surface by the silk-screen process.[7] Subminiature audio amplifier assemblies of this type have wide application in hearing aids, electronic stethoscopes, citizen radio transceivers, etc.

4. Counter

In order that they may be easily fabricated, some subminiature counters consist partly of printed circuits resembling a conventional terminal board structure with fired-on silver conductors instead of turret lugs and hook-up wire. Each decade is mounted on a single steatite plate composed of four binary counters connected in such a manner as to count tens. A unique structure that has been suggested for a single binary counter assembly is composed of a cylinder of high dielectric constant ceramic and a T-3 subminiature double triode that is inserted into the center of the cylinder. Resistors, capacitors, and wiring may all be printed, with the cylinder itself functioning as the capacitor dielectric medium. A complete binary counter could be designed to occupy a cylindrical space with an outside diameter of ½ in. and a length of 1½ in.

FIG. 2. Front and back views of a monoplanar printed, high-gain hearing-aid audio amplifier (by courtesy of Centralab Division of Globe-Union Inc.).

An interesting plug-in ring counter employing subminiature thyra-

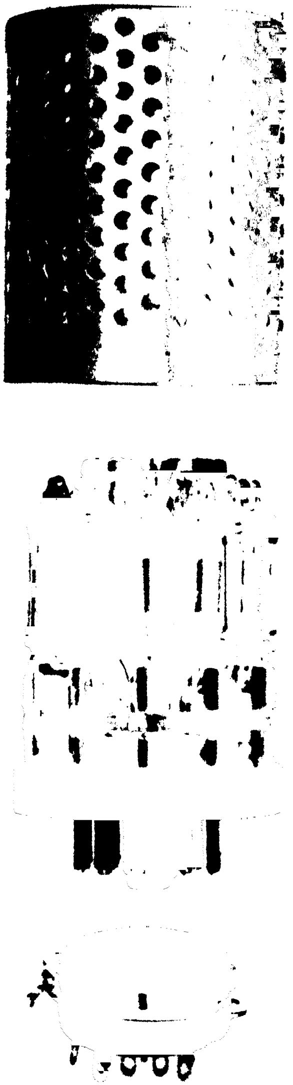


FIG 3 Subminiature decade ring counter using thyratrons (by courtesy of Sylvania Electric Co)

trons is illustrated in Fig. 3.[8] The chassis of each stage of the counter is a printed steatite plate on which are mounted the component parts and the thyratron. These plates are then stacked radially to form a complete decade that will fit into a can 2-in. in diameter and 2½ in. high. In this small package are housed 11 tubes, 78 resistors, and 23 capacitors. Perforations are provided in the can cover for ventilation.

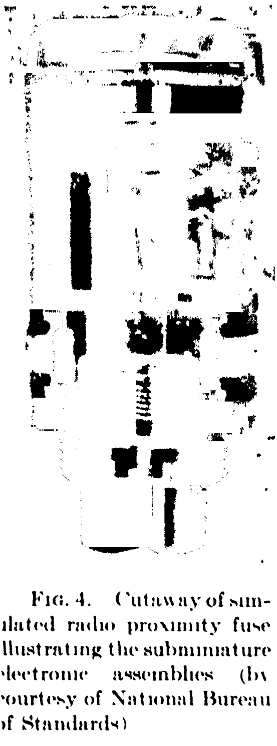


FIG. 4. Cutaway of simulated radio proximity fuse illustrating the subminiature electronic assemblies (by courtesy of National Bureau of Standards)

5. *Proximity Fuze*

No discussion of subminiature devices is complete without mention of the radio proximity fuze. Figure 4 illustrates a simulated cut-away model of a fuze for a mortar shell. A large portion of the circuit has been applied to steatite plates by means of the stencil screen process.[7]

6. *Potted Amplifier*[9]

It would be very desirable if a group of components could be merely interconnected and potted. Unfortunately this is not at all practical. When fabricating a potted electronic assembly some means must be provided for supporting the components while interconnecting them and to prevent short circuits during the potting process. One very ingenious technique for accomplishing this is illustrated in Fig. 5. The chassis is an L-shaped plastic channel into which are molded parallel wire conductors. Connections are made by simply removing the plastic surrounding the conductors at appropriate places. The wires emerging from one end of the channel are bent to provide pins spaced to fit a miniature seven-pin socket. The parallel wires may be cut anywhere along their length to provide an interconnection between components not terminated in a pin. The vacuum tubes are provided with resilient jackets so that they may resist the strains of potting. A spacer is placed over the pins to position them before potting. The potting compound used in this assembly is a special resilient body.

7. *I-F Amplifiers*

As wide-band radar type i-f amplifiers have received more attention from subminiaturization engineers than any other type of circuit, the art is at a high state of development. Each of the three techniques described below has something worthy of recommendation.[10] Thus the Model 2 design attempts to provide an assembly that uses the fewest printed circuits and whose fabrication techniques are more nearly conventional. The intent of the P.C. 4 design is the exact opposite. There, emphasis is on the maximum possible use of printed circuits. The model 5 design is aimed toward the smallest practical amplifier design.

a. NBS Model 2. The Model 2 amplifier[5] is centered at 60-mc with a 10-mc bandwidth and a gain in excess of 95 db. It is conservatively

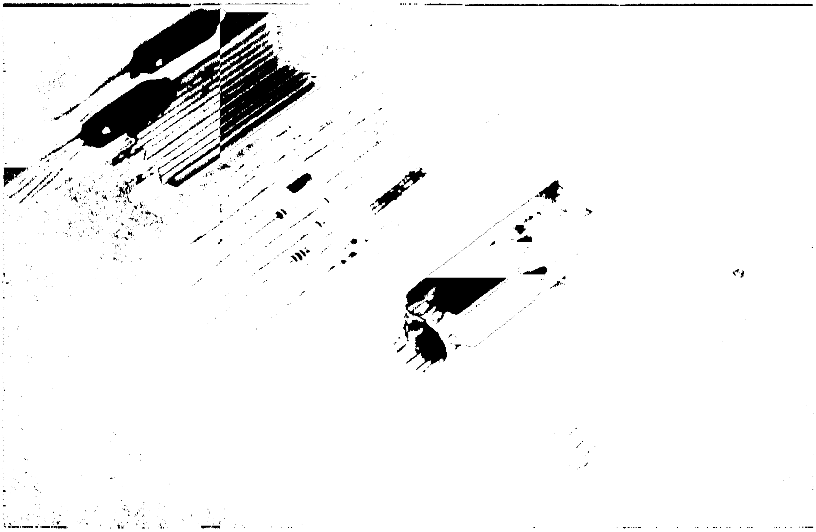


FIG. 5. A novel method for assembling the piece components of a potted amplifier (by courtesy of Melpar, Inc.).

designed and employs more tubes than are required to achieve the desired performance. It consists of four staggered doubles, a diode detector, a video amplifier, a cathode follower, and a tuning indicator diode. The interstage networks are bifilar-wound inductors. The resistors may be either miniature, cracked-carbon types, or conventional composition types. The steatite inductor forms function as resistor terminal boards with silvered ends and holes into which the resistors may be slipped. With the exception of two pieces of hook-up wire, all interconnections are furnished by the tube leads, resistor leads, choke leads and the wiring patterns on the inductor forms. The capacitors are fabricated from high dielectric constant ceramic tubing. All insulating materials and solder are high-temperature varieties. The case is hermetically sealed and filled with an inert gas. The overall dimensions are approximately 11 in by $\frac{9}{16}$ in. by $1\frac{3}{4}$ in. Figure 6 illustrates a breakdown of this amplifier structure.

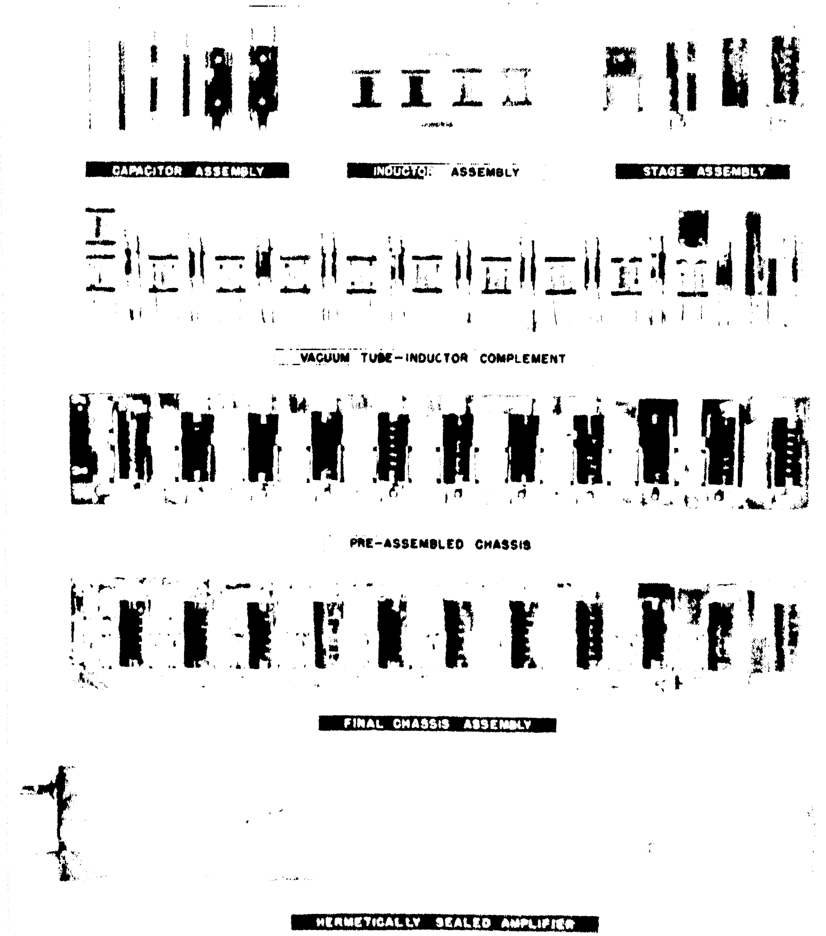


Fig. 6 NBS Model 2 subminiature radar type i-f amplifier and its subassemblies (by courtesy of National Bureau of Standards).

b. *NBS Model PC 4* With the exception of the wire-wound inductors and chokes, this amplifier assembly, Fig 7, is composed entirely of printed circuits The type of printed circuit is unusual, differing from the monoplanar variety usually encountered [11] The design is based on a basic stage module roughly cylindrical in shape, approximately $\frac{1}{2}$ in in diameter and 2 in. long. The basic element is a high dielectric constant ceramic cylinder with a $\frac{1}{2}$ in. outside diameter, $1\frac{1}{2}$ in. long with a hole in the center large enough to contain a T-3 subminiature vacuum tube. The inner surface of the cylinder is completely silvered and the outside surface is silvered in patches. The patches function as the high-potential electrodes of bypass capacitors, the inner silvered coating functions as the common ground electrode, and the cylinder functions as the dielectric medium. The interstage inductor is wound of wire on a steatite form and is placed at the base of the vacuum tube, coaxial with the vacuum tube and the ceramic cylinder. Not only are the capacitors printed on the cylinder, but the silvered interconnections and the printed tape resistors are also located there. A vitreous enamel overcoat insulates the printing from the sheet-metal support structure. A metallized steatite cylinder surrounds and shields the interstage inductor. Schematically, this assembly composed of the 11 modules illustrated, is almost identical with the Model 2 assembly previously described. The Model PC 4 is approximately $6\frac{3}{4}$ in. by $2\frac{5}{16}$ in. by $\frac{11}{16}$ in.

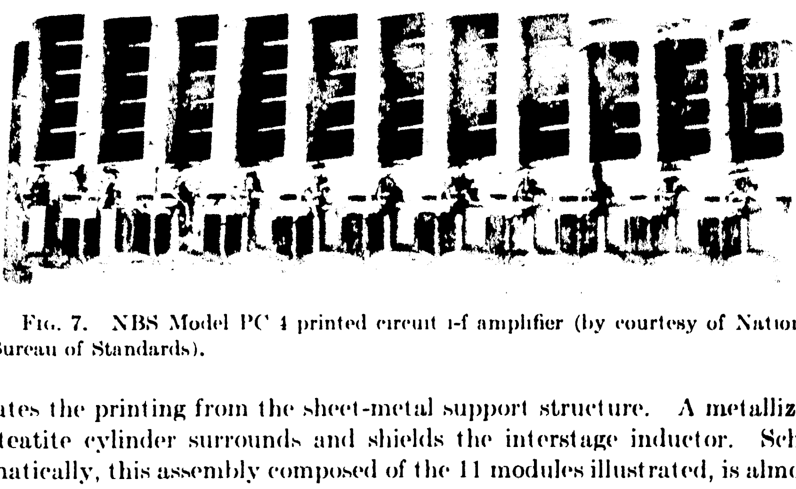


FIG. 7. NBS Model PC 4 printed circuit i-f amplifier (by courtesy of National Bureau of Standards).

c. NBS Model 5. The NBS Model 5 amplifier illustrated in Fig. 8 is functionally similar to the two previously described.[10] It is considered to be the smallest practical assembly of this general type using T-3 tubes. Each stage is mounted on its own light-gage sheet-metal chassis, and the inductors are mounted axially with the vacuum tubes at their bases. The inductor forms support the resistors in a manner analogous to that in the Model 2. The stages are placed side by side and slipped into a metal case. Spring fingers on each stage chassis make contact with the metal case, completely shielding each stage. This is one of the features that contributes toward the high degree of stability of this design despite its small size The amplifier is $4\frac{3}{4}$ in by $2\frac{7}{8}$ in by $\frac{5}{8}$ in when completely cased

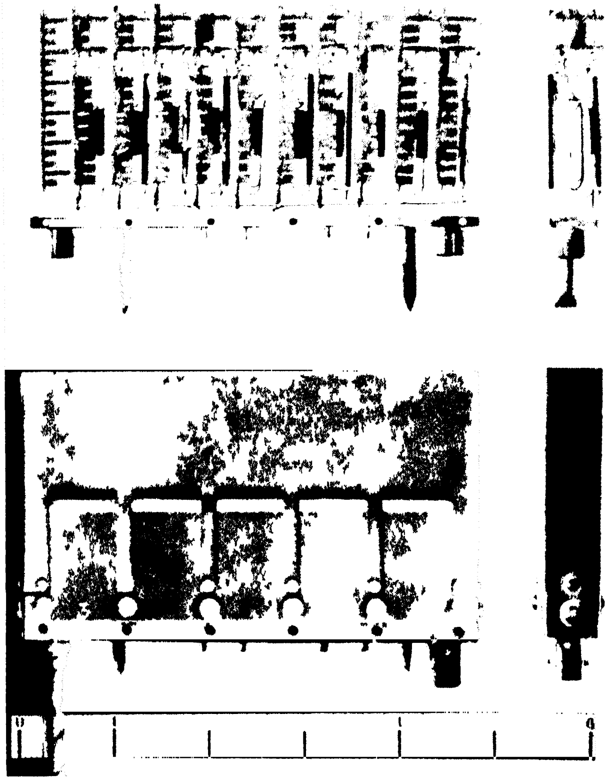


FIG 8 NBS Model 5 subminiature i-f amplifier assembly (by courtesy of National Bureau of Standards)

VI COMPONENTS AND MATERIALS

1 *Insulating Materials*

Although battery-operated subminiaturized equipment and other electronic equipment that is not subjected to high operating temperatures may use conventional insulation electronic assemblies that dissipate considerable power and operate at high temperatures require special insulating materials In the presence of moderate potential stresses and inert surroundings the rate of deterioration of insulating materials is primarily a function of operating temperature Thus the highest temperature at which an insulating material may be used is determined by the life required of that material. The useful life commonly desired in military electronic equipment is 5000 hours.

When used as binders for asbestos, mica, or glass, some grades of phenolic resin will function satisfactorily for more than 500 hours at 200°C. Glass cloth laminates bonded by these materials make very fine terminal boards. Polystyrene and lucite find little application as high-temperature insulating mediums because their heat distortion points are approximately 85°C and 88°C, respectively.

The silicone resins may be used up to temperatures of 250°C. The silicones are available in a number of physical forms such as oil, varnish, elastic, grease, and molding powder. Among the most useful of the silicone products to the subminiaturization engineer are the varnishes. These varnishes may be used as impregnants for transformer windings, glass tapes, r-f inductor windings, and high-temperature insulating films and as adhesives. As a general rule, the harder silicone varnishes are less flexible than the softer varnishes

Teflon (polytetrafluoroethylene) passes directly from the solid to the gaseous state at 435°C. Because adhesives do not adhere readily to Teflon, it is useful in the design of impregnated assembly jigs and fixtures. Its resistance to chemical attack makes it a useful high-temperature gasket material.

Vitreous enamels are highly useful for providing some metals with a high-temperature insulating surface and for bonding ceramics. There are many varieties of vitreous enamel with various coefficients of expansion. Generally, a vitreous enamel with an expansion coefficient similar to that of the material it coats will prove most satisfactory.[11] The vitreous enamels usually require firing temperatures in the region of 700°C.

Steatite-like bodies are useful in subminiaturization because of their low dielectric losses, high-temperature characteristics and ease of fabrication on a production basis. They may be used as base materials for printed-circuit assemblies or may be fabricated into coil forms, terminal strips, and component catacombs.

2. *Capacitors*

Both metallized- and impregnated-paper capacitors are available in very small sizes, but unfortunately they may not be used in all subminiature designs because of their susceptibility to deterioration at continuous high temperatures. Although mica capacitors may be used at fairly high temperatures, they are not desirable for most designs since their shape factor may not be altered easily. Their cost, coupled with their potential scarcity in time of war, makes them a poor choice for military electronic equipment.

Thin Teflon sheets may be cut from cylinders of Teflon for use as dielectric film in capacitors manufactured somewhat as paper capacitors are made. At present, sheets produced in this way are neither sufficiently thin nor of sufficiently high quality to permit Teflon capacitors to compete in size with ceramic capacitors. Efforts are being made to produce satisfactory thin Teflon films by other methods, and it is likely that a small high-temperature Teflon dielectric capacitor will soon be available.

One of the mica capacitor substitutes developed during World War II was a glass film dielectric capacitor. A large glass manufacturer has successfully produced uniform glass ribbon in thicknesses less than 0.001 in. Despite its extreme thinness, its great flexibility permits handling without too much difficulty. The capacitors are fabricated by laying down alternate layers of glass ribbon and metal foil (usually aluminum) and then fusing the entire structure into a solid mass. These capacitors are available in limited quantities.

Another wartime mica capacitor substitute was the vitreous enamel capacitor developed for the Signal Corps.[12] Essentially, this capacitor is made by spraying alternate layers of vitreous enamel and silver paint through masks and then firing them into a solid laminated structure. A modification of the process uses a silk screen and a squeegee. This process offers many distinct advantages. For example, the shape factor of the capacitors is readily controlled. The capacitors can, if desired, be fabricated into one capacitor block together with circuit wiring and can also function as the assembly chassis. The raw materials required are not expected to be scarce strategic materials. Vitreous enamel capacitors have very low losses and low temperature coefficients. The low dielectric constant, however, makes it impossible to obtain high capacity in small volume.

With the exception of the electrical ceramic specialist, few engineers are familiar with the properties of the more common high dielectric constant ceramic bodies. Often what is termed a 1000-$\mu\mu$f ceramic bypass capacitor is used without the user being aware that the capacity may vary four-to-one over the operating temperature range of the equipment.

The following is a review of high dielectric constant ceramic properties. Ceramic capacitors exhibit a constant temperature coefficient only over restricted temperature ranges. As a general rule, the lower the dielectric constant, the wider is the temperature range over which the temperature coefficient is constant. Capacitors that are used for temperature compensation purposes and those with zero temperature coefficients are usually found only in the lower capacity values because moderate dielectric constant bodies must be used.

The rutile form of titanium dioxide has a dielectric constant of 95 to 105 at room temperature, with a fairly large negative temperature coefficient. The addition of other materials, e.g., magnesium titanate or zirconium dioxide, makes the temperature coefficient less negative and reduces the dielectric constant.

When magnesium oxide is added to titanium dioxide, the resultant has a positive temperature coefficient and a dielectric constant of 13 to 17.

The addition of calcium oxide to titanium dioxide yields a temperature coefficient slightly more negative than titanium dioxide alone and a dielectric constant of 150 to 175

The addition of strontium oxide to titanium dioxide yields a negative coefficient that decreases as the ratio of titanium dioxide to strontium oxide is increased. The dielectric constants range from 225 to 250.

The combination of equal parts titanium dioxide and barium oxide produces some rather unusual characteristics. A graph of dielectric constant plotted against temperature gives a peaked curve with minor peaks on the lower temperature slope. The major peak occurs between 95°C and 120°C in the vicinity of the Curie point with the closest minor peak occurring between −15°C and +10°C. At the peaks, the dielectric constant can be as high as 12,000. The Curie point for ferroelectrics is analogous to that which occurs in ferromagnetic phenomena. At this point, the crystal structure of the barium titanate changes from the stable low temperature tetragonal form to the cubic high temperature stable form. The loss factor of the dielectric also exhibits peaks at the temperature where the dielectric constant is maximum. The major effect of high electric field strength is to increase the dielectric constant and loss factors. This effect is magnified at the peaks. As the ratio of titanium dioxide to barium oxide increases, the dielectric constant starts to drop and the peaks begin to disappear. The temperature coefficient becomes less and less positive, passes through zero, and then becomes negative[14] (Fig. 9).

Although there are a large number of bodies with dielectric constants above 500 containing ingredients other than barium titanate, most of them have one property in common: when heated above 85°C for a few minutes and then reduced to a lower temperature, they exhibit a higher dielectric constant than they did originally. Upon aging, the magnitude of the dielectric constant decays exponentially. Typical curves are illustrated in Fig. 10.[15] It is common for the heat of soldering to change the capacity considerably. This phenomenon appears to repeat itself indefinitely.

$BaTiO_3$, one of the more useful high dielectric constant bodies, unfortunately displays high leakage characteristics at temperatures approaching 200°C. This leakage is especially marked when the dielectric body is very thin (0.005 in. to 0.010 in.) and increases very rapidly as the d-c polarizing voltage becomes greater

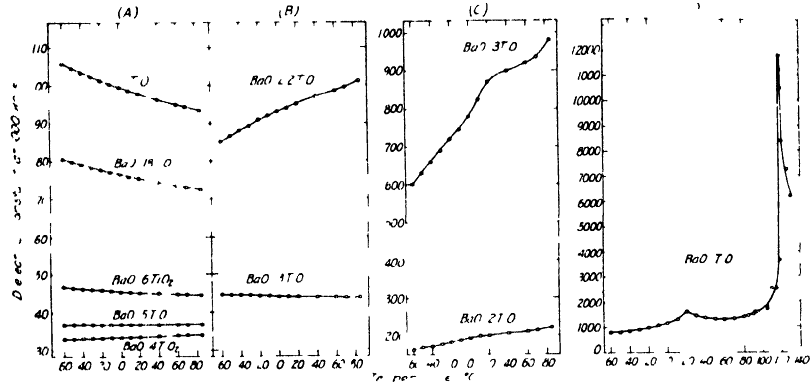


FIG 9 Variation of dielectric constant with temperature at 1000 kc/s, for specimens with compositions in the binary system, TiO_2-$BaTiO_3$ (by courtesy of National Bureau of Standards)

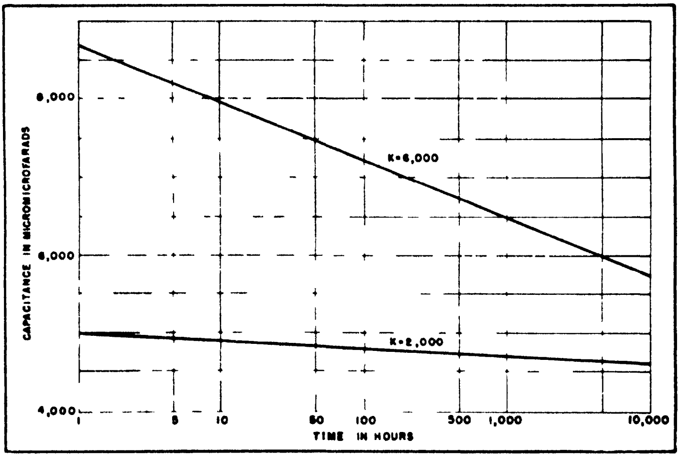


FIG 10 Aging curve of two high K ceramic dielectric materials

The well-known tendency of ordinary aluminum electrolytic capacitors to lose capacity rapidly at low temperatures has barred their use from most military equipment. However, recently developed capacitors having sintered tantalum electrodes are free from this defect.

A subminiature ganged variable capacitor assembly (365 $\mu\mu$f per section) has been designed by one company. The electrodes are nested nickel cylinders insulated from one another by hard thin coatings of nickel oxide. The very thin oxide skin that can be formed on the nickel is the clue to the high capacity that can be obtained with a small pair of cylinders.

3. Conductive Paint

One of the more useful aids in subminiaturization is silver paint. The formulations and methods for using such paints are described in the literature.[4] Essentially, they consist of three ingredients: finely divided silver, powdered low-melting-point glass, and a volatile vehicle. Silver is used as the conducting medium because its corrosion products are also good electrical conductors. The glass functions as a binder. If the percentage of glass is high, adhesion is improved, but the electrical resistance of the fired coating increases, and soldering to the silvered surface becomes difficult. If the percentage of glass is low, the resistance is low and the silvered surface may be soldered with ease, but adhesion is poor. The amount of volatile vehicle adjusts the viscosity for spraying, dipping, painting, etc. It is advisable to buy the silver paints already mixed from commercial sources.

4. Resistors

Although various new types of fixed resistors have appeared in subminiature form, there is a scarcity of test information that may be used to evaluate their relative performance with conventional types.

Resistors formed by cracking carbon on a smooth ceramic rod show much promise if the problem of making electrical contact to the resistance film can be adequately solved.

Silk screen printed circuit resistors for the most part use phenolic binders and are therefor limited to moderate temperature operation.

The printed circuit i-f amplifier previously described uses tape resistors.[11] These tapes are asbestos paper strips with an uncured, silicone-bound resistance film on one side. Lengths of this adhesive resistance tape are laid down on the printed plate. Heating the entire plate simultaneously cures and permanently secures the resistance element.

A number of companies manufacture hearing-aid type potentiometers. These utilize the knob as an outside case. A few versions of conventionally shaped subminiature potentiometers are available also. All of these types employ phenolic-bound resistance elements.

A high temperature potentiometer element ½ in. in diameter has been designed using a tape resistor. The base is a short glass cylinder with fired-on silver commutator bars spaced around the circumference, each parallel to the axis of the cylinder. A tape resistor is laid over half of the commutator. A brush wipes the exposed portion of the commutator, thereby making contact to the resistor element. An evaporated or fired metal film could be substituted for the tape.

5. R-F Inductors

Small toroidal inductors wound on magnetic cores find application in some subminiature designs. Because they have very little leakage inductance, they may be packaged close together with little or no shielding between them. They may be coupled either capacitively or through low-impedance links.

Bifilar unity-coupled inductors find wide application as compact selective interstage elements because their use minimizes the number of required parts and simplifies the assembly wiring.[5]

Various magnetic ferrite bodies are available with Curie temperatures up to 260°C.[16] Although the permeabilities of the higher-temperature ferrites are lower than the lower-temperature varieties, in many inductor designs they yield higher Q's than can be obtained with powdered iron

6. Transformers

Choice of materials is the major consideration in the design of reactors and transformers with small physical dimensions. The increased internal heating and the higher ambient temperatures at which subminiature transformers are usually made to operate demand the use of materials which will stand operating temperatures as high as 200°C. In recent years, rapid progress in subminiaturization has followed the development of high-temperature wire insulations, new steels for core materials, new insulation materials, and high-temperature impregnants.

The choice of wire for windings must be guided by insulation thickness as well as by temperature and voltage requirements. To calculate the space utilization factor of various wire types a square lay will be assumed. Square cross section wire with a side D ($D = 2R$) utilizes the space 100%. The space utilization of round wire with a radius R equals the ratio of the areas of the round wire cross section to the square wire cross section, multiplied by the factor one hundred.

$$\frac{\pi R^2}{(2R)^2} \times 100 = 25\pi$$

The per cent of copper in the cross section of an insulated round wire is as the ratio of the squares of the copper diameter to the overall diameter, multiplied by 25π. Therefore:

$$\text{Per cent space factor} = 25\pi\left(\frac{\text{copper diameter}}{\text{overall wire diameter}}\right)^2$$

A graphical comparison of the per cent space utilization of various insulated wire types is presented in Fig. 11.

Glass fiber, silicone-bonded, single- or double-wrapped wire has good temperature and voltage characteristics, but because of its thickness has a poor space factor - similar to cotton-wrapped wire Its insulation will unravel when heated at the curing temperature of a high-temperature impregnant.

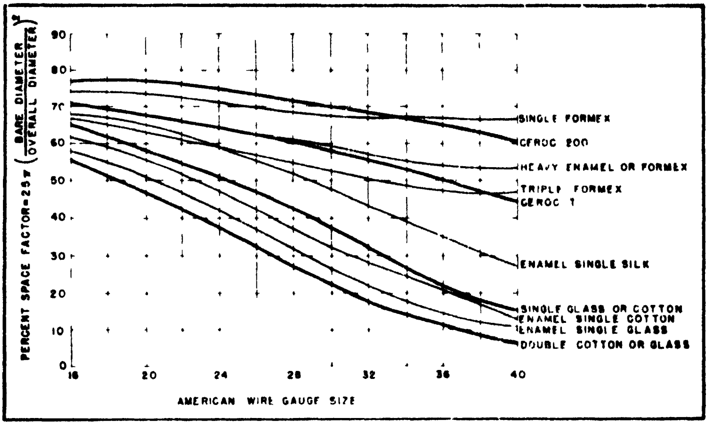


FIG. 11. Space factor of various magnet wires.

Ceramic-insulated wire has an excellent space factor similar to wire with single Formex insulation and will withstand high temperatures. This insulation is a "flexible ceramic" deposited electrically from a suspension and must be handled carefully to avoid abrasion. However, it provides the best space factor for high-temperature windings. Its voltage characteristic is greatly improved by silicone impregnation.

A ceramic insulated wire with an outer overcoat of Teflon is also available. It has a space factor similar to that of a wire coated with heavy enamel- considerably better than glass-wrapped—and is quite flexible. Its voltage rating is higher for equal thicknesses of insulation than that of any of the other wires without impregnation. It is not easily abraded and is very slippery. As there is no known material which will adhere to it readily, its smooth surface causes some difficulty in winding coils and securing ends.

Teflon's high-voltage rating permits its use in random windings as well as layer windings. It is impervious to moisture and acids and is not readily damaged by abrasion in a coil as a result of thermal expansion. It is therefore not necessary to impregnate most types of coils wound with Teflon-insulated wire. When an impregnant is used, it is mainly for mechanical reasons. Impregnants do not actually adhere to the teflon.

Silicone-varnished wire has shown good space factors and electrical characteristics. It has been found, however, that when a transformer using this wire is vacuum impregnated with silicone varnish, the pressure causes the softened insulation to be squeezed out of the winding. This type of wire is still under development.

Where grain-oriented steels are not required, shell type cores are preferred to type "C" cores because only one winding assembly is needed. In the case of a reactor for a power-supply filter, no reduction in size can be obtained by using grain-oriented steel if a large d-c current is present. With a d-c saturation of 120 or more ampere-turns per inch of mean magnetic path, the incremental permeability of a core made of interleaved laminations of 4% hot-rolled silicone steel is higher than that of a type "C" core with the same magnetic path length in grain-oriented steel.[17]

For power transformers, the most suitable core material from the standpoint of minimum size is one which has the highest saturation point without excessive core losses and excitation. Type "C" cores of Westinghouse C-95 Hipersil (0.005-in thick wound strip), a grain-oriented cold-rolled silicon steel, have been found suitable for power frequencies of 400 to 800 cps. For lower-power frequencies, type C97 Hipersil (0.013-in. thick strip) is useful.[18]

Forms or core tubes can be made from the following materials:

(*a*) Mica splittings coated with partially cured silicone varnish.

(*b*) Small overlapped mica scales bonded to fiber glass with silicone varnish.

(*c*) Asbestos paper, available in sheet or tape form in various thicknesses which can be bonded with silicone varnish in the fabrication of core tubes.

Sheets of Teflon, silicone-coated fiber glass, and asbestos paper are available for interlayer insulation.

The limits of transformer subminiaturization are determined by:

(*a*) The minimum wire size which can be handled in winding.

(*b*) The high copper loss of small-gauge wire.

(*c*) Saturation effects in cores of small cross section.

(*d*) The thickness of a particular piece of insulation and the creepage distance, which are constant for a given voltage to ground, regardless of the size of the winding.

The cost of subminiaturized transformers will be considerably higher than those of conventional size, but since transformers and reactors constitute the larger components in most circuits, a substantial reduction in their size and weight can effect a very desirable overall reduction in the weight and size of the whole chassis.

7. *Batteries*

A possible replacement for the lead-acid storage cell is a new cell utilizing silver and zinc as the active materials in an alkaline electrolyte. These cells, it is claimed,[19] are lighter and smaller than other storage battery cells of the same capacity. The number of charging cycles is greater than that of the normal lead-acid type and compares favorably with the ferro-nickel or cadmium-nickel batteries.

A new dry cell similar in general to the Leclanche type cell, but differing from it in certain major respects, has been marketed recently. No metal oxides are used as depolarizers, this effect being secured by use of highly active, oxygen-absorbing carbon electrodes. A stable gel paste, supported by a fibrous matte, immobilizes the required electrolyte which, to all intents and purposes, is regenerated as used and remains unchanged throughout the service life of the cell. The net total of these reactions is simply that of "burning" the pure zinc anode in oxygen from the air through the active-carbon cathode The carbon electrode is on the outer surface of the cell, where it is best exposed to the air. The high efficiencies of both volume and weight, coupled with a remarkably flat discharge curve, combine to make up a very useful primary battery.

The Ruben cell,[20] a new alkaline dry cell which makes use of the electrochemical system $Zn\ Zn(OH)_2$(solid) KOH(Aq)HgO(solid), has been used quite successfully by the military services for some years. The nominal voltage of this cell is 1.34, but it can be used interchangeably with the conventional dry cell. Its discharge curves are unusually flat. The service obtainable from the Ruben cell under high current drains is from 4 to 7 times that of conventional dry cells of equivalent volume. Shelf life is unusually long.

VII. OUTSTANDING PROBLEMS

Far from being completely solved, there still remain a number of vital basic problems associated with subminiaturization.

1. The present size of power transformer-rectifier-filter power supplies is all out of proportion to the size of the subminiature electronic assemblies that they power.
2. In the portable equipment field there is still need for a cheap, light, small, high-drain battery with long shelf life.
3. With electronic equipment being required to operate at ever higher temperatures, a new basic approach to the cooling problem is badly needed.
4. Flexible automatic and semiautomatic methods requiring only a few unskilled assembly personnel for the manufacture of complex subminiature assemblies must be taken from the laboratory and sold to industry.
5. Circuitry that will eliminate large components and reduce the number of components required to perform particular functions is needed.[2]

VIII. CONCLUSIONS

Experience with subminiaturization techniques has demonstrated the validity of the following generalizations:

1. A subminiaturized design requires more development engineering time than a conventional design because of the additional problems inherent in extremely compact assemblies.
2. Rarely is an initial subminiature design satisfactory; it is usually necessary to construct a second and sometimes a third model before a final product is achieved. It is seldom possible to "cut corners" by making all the designs except the final one of the paper variety. It is only by actually constructing the initial designs that assembly and production difficulties are uncovered and means for overcoming them conceived. The engineer must also exercise care that his design is not made impractical by neglecting to provide for ease of manufacture, alignment, and repair.
3. Greater rigidity and shock resistance can be expected of subminiature assemblies because of the smaller masses involved.
4. Because they lend themselves to plug-in package designs, subminiature assemblies can greatly ease equipment maintenance problems.
5. It appears, from techniques developed thus far, that production costs of equipment utilizing subminiature designs will compete favorably with costs of conventional electronic manufacturing practices.

REFERENCES

1. U. S. Department of Commerce, Bureau of Census. Census of Manufactures. No. MC36D, 1947.

2. Sultzer, P. G. Circuit techniques for miniaturization. *Electronics*, **22**, No. 8, 98–99 (1949).
3. Cuming, W. R. Plastic embedded circuits. *Electronics*, **23**, No. 6, 66–69 (1950).
4. Brunetti, C., and Curtis, R. W. Printed circuit techniques. *Natl. Bur. Standards* (*U.S.*), Circular No. 468.
5. Natl. Bur. Standards (U.S.). Final Report, Electronic Miniaturization. #PB100-949 Office of Technical Services, U. S. Dept. of Commerce.
6. Natl. Bur. Standards (U.S.) Development of the National Bureau of Standards Casting Resin. NBS Circular 493.
7. Brunetti, C., and Khouri, A. S. Printed electronic circuits. *Electronics*, **19**, No. 4, 104–108 (1946).
8. Meinheir, C. E., and Snyder, W. Electronic counter and divider circuits. *The Sylvania Technologist*, **1**, No. 3 (1949).
9. Tuller, W. G. Potted subassemblies for subminiature equipment. *Electronics*, **22**, No. 9. 104–105 (1949).
10. Shapiro, G., Henry, R. L., and Seal, R. K-F. New techniques for electronic miniaturization. *Proc. Inst. Radio Engrs.*, **38**, No 10 (1950).
11. Natl. Bur. Standards (U.S.) Final Report, Printed Circuits. #PB100950 Office of Technical Services, U. S. Dept. of Commerce.
12. Bradford, C. I., Weller, B. I., and McNeight, S. A. Printed vitreous enamel components. *Electronics*, **20**, No. 12, 106–108 (1947)
13. Titanium Alloy Mfg. Co. High dielectric constant titania and titanate ceramics.
14. Bunting, E. N., Shelton, G. R., and Creamer, A. S. Properties of barium-strontium titanate dielectrics. *Nat. Bur. Standards* (*U.S.*), Research Paper No. 1776.
15. Marks, B. H. Ceramic dielectric materials *Electronics*, **21**, No. 8, 116–120 (1948).
16. Snyder, C. L., Alberts-Shoenberg, E., and Goldsmith, H A Magnetic ferrite core materials for high frequencies. *Elec. Mfg.*, **44**, No 12 (1948).
17. Bell Telephone Laboratories. High Operating Temperature Transformers. Final Report, A.M.C. Contract W-33-038-AC-13939, June, 1948.
18. Lee, Ruben. Electronic Transformers and Circuits. John Wiley and Sons, 1947.
19. Literature of the Yardney Electric Corporation, New York.
20. Friedman, M., and McCauley, C. E. The ruben cell –A new primary alkaline battery. Paper presented at the Electrochemical Society Meeting. Oct. 15–18, 1947.

Principles of Pulse Code Modulation

H. F. MAYER

School of Electrical Engineering, Cornell University, Ithaca, New York

CONTENTS

		Page
I.	Introduction	221
II.	Short Survey of Noise-Cleaning Methods	222
III.	The Sampling Theorem	226
IV.	Quantization	229
V	Encoding	233
VI	Principal Operations at the Transmitting End	. 237
VII.	Principal Operations at the Receiving End	243
VIII	Fidelity in PCM Transmission	247
IX	Rate of Transmission	.. 256
	References	.. 260

I. INTRODUCTION

The growth of modern communication networks towards a global network involves longer and longer communication circuits. These long circuits consist in general of a very large number of individual repeater sections or radio links. With present techniques, a long distance circuit of 3000 miles in a coaxial cable consists of about 500 repeater sections. The same circuit, built up as a microwave connection, would include about 100 individual radio links in tandem.

Each section or link is subject to noise and noise cumulates with the individual links. For a given quality of overall transmission, the requirements on each link become more severe as the length of the circuit increases. Noise therefore is a very serious problem in long distance communication. The ultimate cost of the very long circuits for transmission of speech, music, and television will largely depend on whether or not one succeeds in finding simple and cheap means to combat circuit noise.

One way to attack the noise problem consists in keeping the transmission channel free of *external* noise. For example, one can use shielded pairs in a multicore cable or highly directional antennas in connection with radio links. Such and other means are self-evident in long distance communication. But they can do nothing against *internal* noise, which results from the corpuscular structure of matter, such as resistor noise from the line or tube noise from the amplifiers.

Fortunately, there exists another way to attack the noise problem. Circuit noise can be more or less wiped out by special *noise-cleaning transmission methods.*

These methods are in contrast to ordinary transmission, where the signal is transmitted over the circuit in its original form, as produced for example by the microphone in voice frequency telephony. Characteristic values of the signal at the transmitting end are the signal frequency band B and the signal power S. Similarly, characteristic values of the circuit are channel frequency band and channel noise power. With ordinary transmission, the channel bandwidth and the signal bandwidth are alike. Signal attenuation within each individual section can be compensated by intermediate amplifiers with the effect, that the signal will be found with the same power S at the receiving end of the circuit. But the numerous intermediate repeaters amplify likewise the internal circuit noise which thus cumulates to a noise power N at the receiving end. Here, both signal and noise are offered to the destination, for example by means of a telephone or loudspeaker. Nothing can be done to improve the once obtained signal-to-noise ratio $S:N$. If with a given type of circuit this ratio does not fulfil transmission requirement the only way to improve it consists in increasing the signal power at the transmitting end of the circuit.

Since the beginnings of communication it was strongly believed that noise cannot be separated from a signal once it has entered the transmission channel and is mixed with the signal. It was therefore highly surprising when E. H. Armstrong demonstrated in 1936 that frequency modulation could be used as a noise-cleaning transmission method. Since then, other noise-cleaning methods have been found and developed, the latest and most effective method being pulse code modulation, in short PCM.

This important method of transmission was originally suggested by Reeves[1] in 1939.

II. SHORT SURVEY OF NOISE-CLEANING METHODS

All noise-cleaning methods have certain features in common.

Figure 1*a* represents ordinary transmission, where a signal function $F(t)$ is applied with a power S to a transmission channel of bandwidth B. Due to the receiving amplifier, the signal $F(t)$ will be received with the same power S, but together with the noise power N, which has accumulated in the band B. The result is a certain signal to noise ratio $S:N$.

Figure 1*b* represents a noise-cleaning transmission method. In general it is impossible to transmit the original (or primary) signal function $F(t)$ over a noisy channel and to obtain at the same time a noise-cleaning effect. The primary signal function $F(t)$, comprising the signal band B, has to be transformed into another function, the secondary signal $G(t)$, by means of a modulator M. The secondary signal $G(t)$ will then occupy a broader band $r \cdot B$, and the noise-cleaning effect will largely depend on the expansion factor r. The channel bandwidth therefore, in order to transmit the secondary signal, has also to be expanded from B to $r \cdot B$.

For comparison, one can assume that the secondary signal $G(t)$ is applied with the same signal power S to the same physical circuit as in case (*a*). Assuming furthermore equal noise power per cycle, as is the practical case with internal noise, an increased noise power $r \cdot N$ will be found together with the signal power S at the output terminals of the expanded transmission channel. The necessity to use a broader transmission channel results in an even worse signal to noise ratio $S:rN$ at the receiving end of the channel. So far the noise-cleaning method has only disadvantages: an additional modulator, a broader and therefore more expensive channel, and a lower signal-noise ratio at the receiving end of the channel.

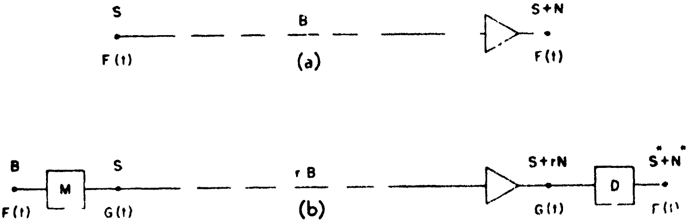


FIG. 1 Ordinary and noise-cleaning method

The payoff comes with the demodulator D, Fig. 1*b*, which transforms the secondary signal $G(t)$ back into the primary signal $F(t)$. In general, the demodulator will perform the inverse operation of the modulator, but has special means to combat noise. These special means and their effectiveness to throw out noise depend largely on the properties of secondary signal function $G(t)$ used for transmission. At the output terminals of the demodulator the original signal function $F(t)$ will be found with a certain signal power S^*. But also some noise will slip through and will be found as noise power N^* in the output signal band B.

One can speak of a noise-cleaning effect, if with a given method

$$S^*/N^* > S/rN$$

But in order that case (*b*) has an advantage over case (*a*), the relation must hold

$$S^*/N^* > S/N$$

The output signal to noise ratios of case (*a*) and (*b*) are related by an equation of the form

$$\frac{S^*}{N^*} = f\left(r, \frac{S}{N}\right)$$

where f is a characteristic function of the special method used for transmission

Under comparable conditions, that is with a given expansion factor r, given signal power S and noise power N/B per cycle bandwidth, different methods result in different signal to noise ratios S^*/N^*. The ideal method would clean out all the noise and thus provide for an infinitely high signal-to-noise ratio. All now existing methods are not ideal in this sense, but pulse code modulation comes nearest to the ideal case. As noise-cleaning has to be paid for by an increased channel bandwidth, pulse code modulation makes best use of this additional bandwidth.

In order to transform the primary signal function $F(t)$ into the secondary signal function $G(t)$ by means of a modulator, a continuous carrier or pulsed carrier can be used. With a continuous carrier $A \sin \phi$, either the amplitude A or the angular displacement ϕ can be varied according to the signal $F(t)$. But only angular modulation, that is *frequency modulation and phase modulation*, will produce a noise-cleaning effect, never amplitude modulation.

With a pulsed carrier, either the pulse amplitude, or the pulse duration or the time distance between successive pulses (pulse position) can be varied with the signal $F(t)$. Principal methods with pulse transmission are therefore pulse-amplitude modulation, pulse-duration modulation, and pulse-position modulation Here again, pulse-amplitude modulation can produce no noise-cleaning effect. But *pulse-duration modulation* and especially *pulse-position modulation* are effective noise-cleaning methods.

It may be seen that in order to perform noise-cleaning, one has to transform the primary signal $F(t)$ into the secondary signal $G(t)$ in such a way that the characteristic variations with the secondary signal occur "in time" rather than "in amplitude," as is the case with the primary signal. By the process of modulation, information is shifted from the amplitude dimension into the time dimension. This permits one to keep the amplitude of the continuous or pulsed carrier constant. As the amplitude of the secondary signal is now no longer a carrier of information, it becomes possible to eliminate amplitude variations, caused by the channel noise, with an amplitude filter at the demodulator. The process of amplitude filtering at the demodulator is an essential feature of all noise-cleaning methods.

With any type of pulse modulation, the free gaps between successive pulses can be used to transmit other pulse series, modulated with other signal functions. This means "time multiplex," whereas multiplex operation with continuous carrier methods is bound to "frequency multiplex." Under certain conditions, especially in connection with microwave transmission, time multiplex has important advantages over frequency multiplex. As a result, noise-cleaning pulse methods are generally used in connection with time multiplex, although noise-cleaning and time multiplex have nothing to do with each other in principle.

In a PCM system, the primary signal function $F(t)$, with its subtle amplitude variations, is also transformed into a secondary signal. This signal is merely a sequence of on-off pulses, or marks and spaces, where neither the pulse amplitude nor the pulse width nor pulse position varies. The only criterion at the receiver is if at a predetermined time instant a pulse occurs or not. This makes it possible, to use not only an amplitude filter, but also a time filter (gate), which opens the receiver only at these predetermined time instants. In spending all available signal power at the transmitter just for 2 discrete amplitude levels, on and off, there is a good chance that the receiver can distinguish between mark and space even in the presence of considerable noise. If a pulse, although considerably misformed by noise, is recognized as such at the receiving end of a link, it can be reshaped and sent out like new over the next link. Reshaping or pulse regeneration can take place at all intermediate repeater points of a long distance circuit with the effect that the transmitted sequence of on-off pulses appears as good as new at the remote end of the circuit. The only condition is that the noise within each link remains below a comparatively large threshold value at which discrimination between mark and space becomes impossible. At the receiver, the demodulator will then retransform the received marks and spaces into the original signal.

Existing PCM systems are designed to transmit simultaneously a large number of speech channels in time multiplex over microwave links.

The most recent PCM-system, described by Meachan and Peterson,[9] is a 96 channel system, where the 96 channels are arranged in 8 groups of 12 channels each. The channels within each group are assembled on a time division basis and the 8 groups are then assembled to one supergroup on a frequency division basis. This article is a very complete description of the most advanced PCM equipment developed by Bell Telephone Laboratories.

It is therefore not the intention of the present article to describe the circuitry of actual PCM systems in detail but to restrict consideration more to basic principles. As neither carrier transmission nor time multiplex are fundamental properties of PCM, the basic principles involved will be discussed in terms of the simplest case, namely one channel and video transmission.

PCM involves certain new and important principles, such as sampling, quantizing, encoding, pulse shaping, and decoding, which will be discussed in the following sections.

III. THE SAMPLING THEOREM

The first basic operation in PCM is "sampling." This operation is based on the *sampling theorem*. A short but comprehensive review of this important theorem was recently given by Shannon.[14]

All communication signals, such as telegraph signals, telephone signals, and television signals, have a characteristic, but finite signal frequency band. The fact that the band is limited has far reaching consequences.

Theorem: A signal function, which contains frequencies only in a limited band of B cps, B also being the highest signal frequency, is completely determined by discrete ordinates at points $1/2B$ seconds apart.

This important theorem may be illustrated by an arrangement shown in Fig. 2. The source at the left produces a continuous signal $F(t)$, limited to the signal band $0 \cdots B$ cps. A switch, which rotates with a constant frequency $f_0 = 1/\tau_0$ connects the source with an ideal low-pass filter, also of bandwidth B, during a very short time τ in regular intervals τ_0. One thus obtains at the switch a series of short pulses $F_1 F_2 \cdots F_n$, all pulses being of equal duration τ and successive pulses τ_0 seconds apart. An ideal amplifier with a linear voltage gain of τ_0/τ follows the low-pass filter.

At the output terminals of this amplifier one obtains a signal function, which is *identical* with $F(t)$ at the source, under the sole condition that the switching frequency is at least twice the highest signal frequency, $f_0 \geqq 2B$, or

$$\tau_0 = 1/2B \tag{1}$$

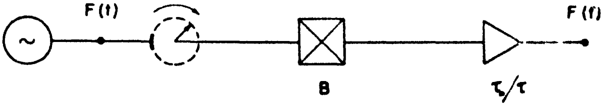
as the limiting case.

The discrete pulses in Fig. 2 are called "samples," more correctly amplitude samples.

Let the nth pulse of amplitude F_n occur at a time $t_n = n \cdot \tau_0$. This pulse, if sent through the low-pass and amplifier in Fig. 2, produces an output pulse

$$V_n = F_n \frac{\sin \frac{\pi}{\tau_0}(t - n\tau_0)}{\frac{\pi}{\tau_0}(t - n\tau_0)} \tag{2}$$

This pulse is shown in Fig. 3. It is centered at a time $n\tau_0$ and will be zero at all other sampling points $t = k\tau_0$, $k \neq n$. Consequently, all other pulses V_k are zero at a time $t = n\tau_0$. There is no correlation between the samples, that is, if one changes the value of one particular sample F_n, such a change will have no effect on any other sample F_k.



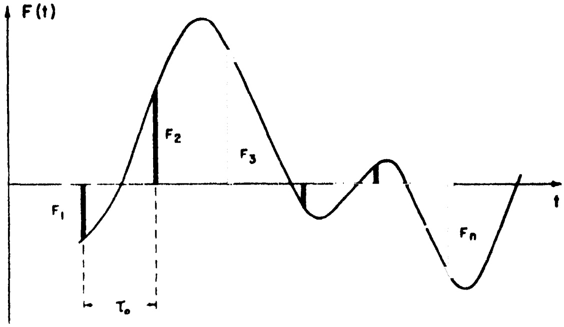


FIG 2 Sampling device.

As the output signal is simply the result of all samples, the only possible representation of the signal as a function of time will be

$$F(t) = \sum_{-\infty}^{+\infty} V_n = \sum_{-\infty}^{+\infty} F_n \cdot \frac{\sin \frac{\pi}{\tau_0}(t - n\tau_0)}{\frac{\pi}{\tau_0}(t - n\tau_0)} \tag{3}$$

A given continuous signal function $F(t)$, limited to a signal band B, can therefore always be considered to consist of a superposition of pulses V_n, all pulses being of equal shape (Fig. 3), but of different amplitudes F_n, centered at time instants $1/2B$ seconds apart, whether it is actually sampled or not.

Thus, if one samples the signal (looks at the source through the rotating switch), one will see only the amplitude of the pulse centered at that time instant, but none of the others, as all other pulses are zero at that time. Due to the limited bandwidth it is sufficient to glance at the signal only at time instants $1/2B$ seconds apart. One then knows everything about the signal and can forget the rest.

Each signal function carries information by the variation of its amplitude. One sees that only the amplitudes at the sampling points carry information, but not the amplitudes between, as these are already determined by the samples. One can consider a sample F_n to be an independent carrier of information, the information carried being the numerical value F_n of the sample.

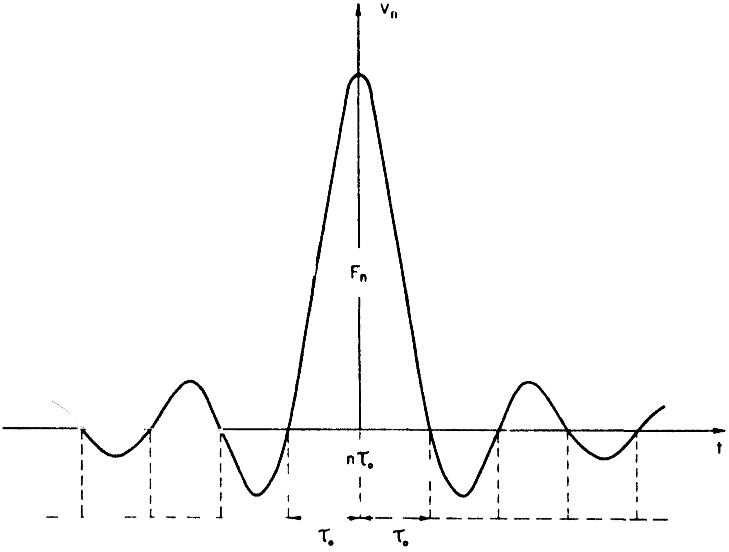


FIG. 3 Output pulse at a time $t = n\tau_0$

During a time interval T seconds one will find $T/\tau_0 = 2BT$ samples. Thus the information content of a signal $F(t)$, limited to a band B and time interval T, is just a set of $2BT$ numerical values.

It may be remembered that in order to describe a signal of bandwidth B and duration T by a Fourier series, one needs a set of discrete frequencies, $1/T$ cycles apart, or BT frequencies in all. Here each frequency sample carries two numerical values, either amplitude and phase shift or sine and cosine term, i.e., $2BT$ numerical values in all. These values are another possible set of numbers which also describe completely the signal under consideration.

To sum up, the problem of transmitting a signal function of signal band B and duration T can always be reduced to the problem of transmitting $2BT$ numerical values (numbers) in one way or another. If one succeeds in reconstructing the discrete samples F_n at the receiver, the original signal function can be reconstructed simply by sending the samples through a low-pass filter of cutoff frequency B and a suitable amplifier, Fig. 2. Between the sampling switch at the transmitter and the low pass at the receiver, one can play with the samples, provided no "information" gets lost

Another consequence of the limited signal bandwidth is this:

If one reconstructs the signal function $F(t)$ by its given discrete samples F_n, the energy E developed in a unit resistance will be

$$E = \int_{-\infty}^{+\infty} F^2(t)dt = \tau_0 \cdot \sum_{-\infty}^{+\infty} F_n^2$$

Hence, if once observes the signal during a finite time interval T, which contains $n = T/\tau_0$ samples, the signal power will be

$$S = \frac{E}{T} = \frac{1}{n}(F_1^2 + F_2^2 \cdots F_n^2) \tag{4}$$

The power of a continuous signal is therefore simply the mean square of its discrete amplitude samples. This power, of course, lies completely inside the signal band B, and no power will be found outside this band.

IV. Quantization

A continuous signal, such as speech, has a continuous range of amplitudes and therefore its samples have a continuous amplitude range. In other words, within the finite amplitude range of the signal one will find an infinite number of amplitude levels. It is now not possible—and also not necessary—to transmit the *exact* amplitudes of the samples. Any human sense, as ultimate receiver, can only detect *finite* intensity differences. If one transmits, for example, just one sample and offers a corresponding sound pulse, Fig. 3, to the ear, it will judge different samples OP to be equal, if P lies within a certain amplitude range α, Fig. 4. It is therefore permissible to represent and to transmit all amplitude levels

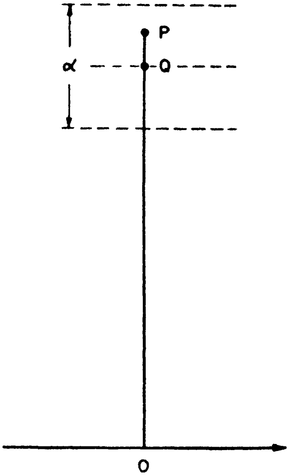


Fig. 4 Quantization step.

within this range by one discrete amplitude level OQ. Hence, transmission of any signal can be achieved with a finite number of discrete amplitude levels.

The existence of a finite number of discrete amplitude levels is a basic condition for PCM. The next step in building up a PCM system would therefore consist in inserting a quantizer between the rotating switch (sampler) and the low-pass B in Fig. 2. Such a quantizer will be described in detail in Section VI, d. The recovered signal at the receiver will then not be quite identical with the signal at the source, but differ somewhat. As the maximum error cannot exceed one half step or "quantum," deviation from fidelity can be kept within tolerable limits by using a sufficiently large number of small steps.[11]

If one considers one particular amplitude level $OQ = F$, Fig. 4, a possible measure of fidelity with respect to this particular amplitude level is the mean square of the Euclidean distance d between OP and OQ

$$d^2 = \overline{[(OP) - (OQ)]^2}$$

Assuming that in the course of time the points P cover the range α with equal density, one obtains

$$d^2 = \frac{1}{\alpha}\int_{-\alpha/2}^{+\alpha/2} x^2 dx = \frac{\alpha^2}{12}$$

or

$$\overline{(OP)^2} = F^2 + \tfrac{1}{12}\alpha^2$$

According to eq. 4, $\overline{(OP)^2}$ is the contribution of all samples within the amplitude range α to the power of the signal without quantization, whereas F^2 is the contribution of the same samples with quantization. Both differ by the quantization noise power $\alpha^2/12$.

If the quantizer has s discrete amplitude levels and accordingly s steps α_i, which need not be equal, and if $p(i)$ is the probability of the level (i), the total quantization noise power will be

$$N_q = \tfrac{1}{12} \cdot \sum_i p(i) \cdot \alpha_i^2$$

With equal steps α, one obtains

$$N_q = \alpha^2/12 \tag{5}$$

In order to distinguish between the signal function with and without quantizing, the latter will be called $F_0(t)$. This is the signal function delivered from the source to the sampler, Fig. 2. The signal function, which is determined by the quantized samples, will be called $F(t)$, and $F(t)$ is the signal received at the remote end of the circuit. Due to quantization, $F(t)$ will differ somewhat from $F_0(t)$. If S_0 is the signal power of $F_0(t)$, as determined by the not quantized samples, and S is the power of the recovered signal $F(t)$, determined by the quantized samples, both powers differ by the quantization noise power, or

$$S_0 = S + N_q$$

In other words, one sends the signal power S_0, but the quantizer splits this power up into the signal power S and the quantization noise power N_q which disturbs the signal. At the receiver, one has always the quantization noise, even with a noiseless transmission channel. This price must be paid for the great improvements which otherwise result from quantization.

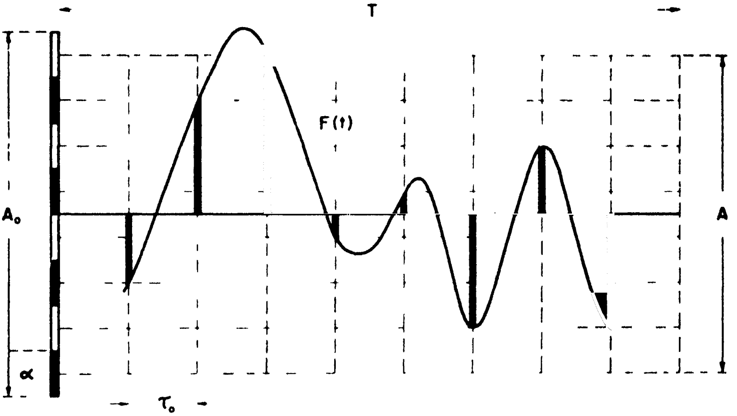


Fig. 5 Quantization with equal steps

As a simple example, the case will be considered where the total amplitude range A_0 of $F_0(t)$ is divided into s equal quanta, Fig. 5. The size of one quantum is then

$$\alpha = A_0/s$$

One sees that the amplitude range of the quantized samples will be

$$A = A_0 - \alpha$$

as the extreme amplitudes are reduced by one-half quantum. Between quantum range A, quantum α and number s of quantum states (amplitude levels) exists the equation

$$s = 1 + \frac{A}{\alpha} \qquad (6)$$

If $F_0(t)$ is sampled and quantized, one has very short pulses at regular time intervals $\tau_0 = 1/2B$. Due to quantization, possible sample amplitudes are only

$$\pm\frac{\alpha}{2},\quad \pm 3\frac{\alpha}{2},\quad \pm 5\frac{\alpha}{2}\ \cdots\ \pm(s-1)\frac{\alpha}{2}$$

If these discrete samples are sent through a low-pass B and a suitable amplifier, Fig. 2, one obtains the continuous signal $F(t)$, Fig. 5.

With a signal function $F_0(t)$, where all amplitudes occur in the course of time with equal probability, the signal power of $F_0(t)$ will be

$$S_0 = \frac{1}{12}A_0^2 = \frac{s^2}{12}\alpha^2$$

In this case also the quantized samples occur with equal probability. The signal power of the recovered signal $F(t)$ is therefore

$$S = \frac{2}{s}[1 + 3^2 + 5^2 + \cdots (s-1)^2]\cdot\frac{\alpha^2}{4}$$

or

$$S = \frac{s^2 - 1}{12}\cdot\alpha^2 \tag{7}$$

As already stated, the quantization noise power is in this case

$$N_q = S_0 - S = \tfrac{1}{12}\alpha^2$$

A measure of fidelity is therefore the signal to noise ratio

$$S/N_q = s^2 - 1 \tag{8}$$

Table I shows how this signal to noise ratio increases with the number s of amplitude levels. Owing to its randomness, quantization noise sounds very much like thermal noise, but is not quite so bad, because no quantization noise will be produced if the circuit is idle.

TABLE I. Quantization signal-noise ratio

s	2	4	8	16	32	64	128
$s^2 - 1$	3	15	63	255	1023	4095	16383
db	4.77	11.76	17.99	24.07	30.10	36.12	42.14

The necessary number s of quantum states at the quantizer depends on the fidelity of transmission desired. Listening tests have shown that 8 or 16 states are just sufficient to obtain good intelligibility of speech, but that quantization noise can easily be detected. Even with the minimum number, 2 states, some intelligibility can be obtained. One considers 32 states to be a minimum for commercial use. The most recent PCM system of toll quality uses 128 states.[9,12]

V. ENCODING

If one intends to keep the original signal function $F_0(t)$ under observation, say during the time interval T in Fig. 5, one must look at all points within the area A_0T, because it is not known in advance where one will find the signal.

With sampling and quantizing, one has to glance at the signal only at time distances $1/2B$ seconds apart, and even then only at certain discrete points at the amplitude scale. This is sufficient to gain all necessary data in order to transmit and reconstruct the signal with sufficiently good fidelity. Only the crossing points of the horizontal and vertical lines in Fig. 5 are necessary observation points. This makes it possible to encode the sampled and quantized primary signal function $F(t)$ into a secondary signal function $G(t)$ by encoding the ordinates of the discrete observation points.

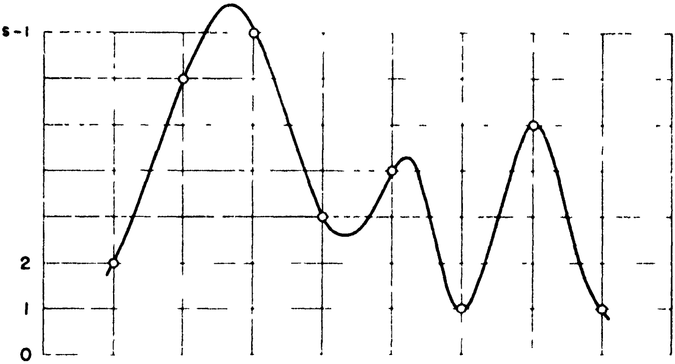


FIG. 6 Primary signal, $s = 8$ states.

For the purpose of encoding, one characterizes the s discrete amplitude levels by integers, the state numbers $0,1,2 \cdots (s-1)$, Fig. 6. The state numbers are then considered to be one-digital numbers of a number system of base s. The process of encoding consists merely in encoding these numbers into another number system of base b.

For the representation of numbers one needs a group of symbols. For example, the decimal number system uses the ten symbols $0,1,2 \cdots 9$, which at the same time represent the first ten integers $0 \cdots 9$ of the number scale. With 2 digits and 10 symbols one can represent 10^2 numbers, with 3 digits 10^3, and so on.

With r digits and a base b (b symbols) one can represent

$$s = b^r \tag{9}$$

numbers, namely the first s integers $0,1,2 \cdots (s-1)$ of the number scale. But admissible values of s are only integer powers of b.

Equation 9 offers the possibility of transforming the primary signal $F(t)$ with s states into the secondary signal $G(t)$ with only b states, at the expense that one needs for each sample of $F(t)$ a group of $r = \log s/\log b$ samples for $G(t)$. These r samples have to be spaced within the time interval $\tau_0 = 1/2B$, Fig. 5. The time interval between the samples of $G(t)$ is therefore reduced to $1/2rB$ seconds. In other words, one has now to transmit $2rB$ samples per second instead of $2B$ samples with $F(t)$ and this requires a broader channel bandwidth rB instead of B. Reduction of states has thus become possible at the expense of an expanded channel frequency band.

According to the choice of b, one can transform $F(t)$ into a large variety of signal functions $G(t)$. The minimum value $b = 2$, or binary number system, offers the greatest advantages in fighting noise. The secondary signal function $G(t)$ consists then only of 2 discrete amplitude levels, on and off. With the binary number system, all numbers are expressed by 2 symbols, 0 and 1, which are called "binary symbols."

If (a) stands either for (0) or (1) the binary notation with r digits

$$a_1 a_2 \cdots a_r$$

represents the number

$$a_1 \cdot 2^0 + a_2 \cdot 2^1 + a_3 \cdot 2^2 + \cdots a_r \cdot 2^{r-1}$$

As an example, the binary notation 101 represents the (absolute) number $1 + 0 + 4 = 5$.*

It is possible, of course, to encode the state numbers also into the ternary number system (base 3) or the quaternary number system (base 4), and so on, but with less and less noise-cleaning effect.

Table II shows as an example the encoding of $F(t)$ in Fig. 6 into binary numbers.

TABLE II. Encoding into binary numbers

State Nos.	2	6	7	3	4	1	5	1
Binary Nos.	010	011	111	110	001	100	101	100

In this case, $s = 8$, one needs a group of 3 binary symbols for each state number of $F(t)$. The binary symbols $(0,1)$ are the state numbers of the secondary signal function $G(t)$. $G(t)$ carries exactly the same message as $F(t)$, but only with 2 states instead of 8 states and with a bandwidth expanded in a ratio 3:1. $G(t)$, of course, can be retransformed into $F(t)$ by the process of decoding, Section VII, c.

---

* Note that binary notations are written "backwards."

$F(t)$ contains $2BT$ samples in a time interval T, and as each sample has to be replaced by a group of

$$r = \log s$$

($\log = \log$ at base 2) binary symbols, the encoded signal will consist of $2BT \cdot \log s$ binary symbols. The message of duration T, either carried by $F(t)$ or $G(t)$, is completely described by this sequence of $2BT \cdot \log s$ binary symbols. This number is used as a measure for the amount of information which can be carried by a signal of duration T, signal band B and s distinguishable discrete amplitude levels. Therefore

$$\text{Amount of information} = 2BT \cdot \log s \text{ bits}$$

or

$$\text{Rate of information } R = 2B \cdot \log s \text{ bits per second} \qquad (10)$$

Each binary symbol (0,1) corresponds by definition to one unit (bit) of information.[14]

If one requests that the message at the receiver is neither compressed nor expanded in time, the rate of information must remain constant with all possible code transformations. That is, if one transforms a signal of bandwidth B_1 and a number s_1 of states into another signal of bandwidth B_2 and s_2 states, the relation must hold

$$R = 2B_1 \log s_1 = 2B_2 \log s_2 \cdots = 2B \log s$$

With number encoding, one can therefore expand or compress the original signal bandwidth B, but only with a corresponding increase or decrease of amplitude states, without impairing the rate of information.

It may be seen by Fig. 4, that the amplitude quantum α acts as a safeguard against channel noise. Noise shifts the point Q in an unpredictable way in an upward or downward direction. But if it stays within the quantum range, that is, if the noise amplitude at this particular sampling point does not exceed $\pm\alpha/2$, one can restore the distorted sample to its original value by quantizing it again at the receiver. Therefore α is a measure of a threshold noise amplitude, or α^2 is a measure of a certain threshold noise power, which can be completely cleaned out due to the finite size of the amplitude quantum.

Signal power S, number of states s, and quantum α are connected by eq. 7.

Hence, if one considers 2 cases, the first case where one encodes the primary signal into a secondary signal with s_1 states and the second case where one encodes it into another secondary signal with s_2 states, but transmits in both cases with the same signal power S, the corresponding values of α^2 will be

$$\alpha_1^2 = \frac{12S}{s_1^2 - 1}, \quad \alpha_2^2 = \frac{12S}{s_2^2 - 1}$$

That is, one obtains maximum discrimination against channel noise if one transmits with a minimum number of states ($s = 2$). Actual PCM systems, where maximum noise cleaning is desired, encode therefore into 2 states. But with a noiseless channel, one could increase the number of states to any large value and thus compress a given signal of bandwidth $B_1 = 2R/\log s_1$ into a bandwidth $B_2 = 2R/\log s_2$ of any desired small value.

A type of noise, which fills out the quantum range α of the secondary signal with equal density, but will never exceed this range, carries a noise power

$$N = \tfrac{1}{12}\alpha^2 \tag{11}$$

Such a type of noise will provide in connection with a signal power S (eq. 8) the space for

$$s = \left(1 + \frac{S}{N}\right)^{1/2}$$

distinguishable amplitude levels on the channel. The capacity of the channel to transmit information without distortion by channel noise is therefore

$$C = 2B \cdot \log s = B \log\left(1 + \frac{S}{N}\right) \text{ bits per second} \tag{12}$$

C is called "channel capacity." That is, if one transmits a quantized signal of power S over a channel of bandwidth B, one can clean out all the channel noise N (by quantizing again at the receiver) provided one feeds the channel at the maximum with C bits per second. It is not possible to transmit more than C bits per second without impairing the signal by the channel noise.[14]

In the case of PCM one needs only $s = 2$ distinguishable states and the critical signal-noise ratio is therefore $(s^2 - 1):1 = 3:1$. All channel noise is completely cleaned out if the noise power will not exceed one-third of the signal power.

But this is only true for the considered type of noise, which has a finite amplitude range and a uniform distribution of amplitudes. Quantization noise for example has such properties but not white noise, and no sharp critical signal-noise ratio exists in this case.

Nevertheless, eq. 12

$$C = B \cdot \log(1 + S/N)$$

has a much deeper significance than one would expect by its above given crude derivation. Shannon[14] has shown, that by properly encoding one can always transmit $B \cdot \log(1 + S/N)$ bits per second with as small an error frequency as desired, but not more. The channel capacity is therefore a sharply defined quantity, not only for quantization noise, but for any type of noise. It assumes, however, this property only in connection with proper encoding.

PCM is the proper encoding method for a type of noise similar to quantization noise but not for white noise, which would involve a much more complicated method of encoding than actually used in PCM.

VI. Principal Operations at the Transmitting End

a. Diagram. Figure 7 shows a simplified diagram of the transmitting end of a single PCM circuit. The source produces a continuous signal $F_0(t)$, limited to the signal frequency band $0 \cdots B$ cps For example, speech currents from a microphone in the range $0 \cdots 4000$ cps may be considered.

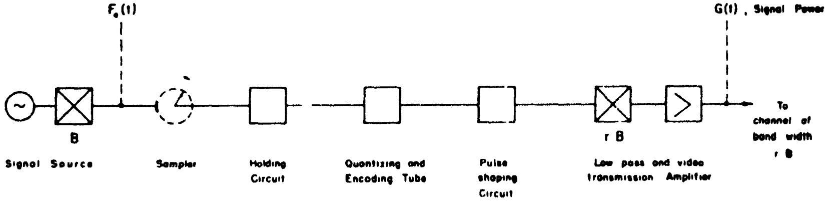

$F_0(t)$
G(t) , Signal Power
B
Signal Source
Sampler
Holding Circuit
Quantizing and Encoding Tube
Pulse shaping Circuit
r B
Low pass and video transmission Amplifier
To channel of band width r B

Fig. 7. Transmitting end of a PCM circuit.

The sampling switch rotates at a frequency $f_0 = 2B$ (8000 cps) and transforms the continuous signal into a series of short pulses, $1/2B$ seconds apart (125 μs.) The duration of each pulse is in the order of a few microseconds.

b. Holding Circuit. These short pulses are now lengthened by a holding circuit. This circuit consists mainly of a capacitor, which, if charged by a sampling pulse, assumes very rapidly a potential equal to the particular pulse amplitude and which holds this potential constant over almost the entire period τ_0. Shortly before the next signal pulse occurs, the capacitor is discharged by a local blanking pulse, synchronized with the sampling switch, and thus becomes ready for the next sample.

c. Coding. Next follows the pulse code modulator, which employs a novel electron beam tube.[10] This ingenious tube performs two operations: it quantizes the amplitudes delivered by the holding circuit to the nearest step of the discrete amplitude scale, and it transforms each quantized amplitude into a group of on-off pulses.

The coding tube, Fig. 8, consists in principle of an electron gun; 2 pairs of deflection plates (X and Y plates), a code masking plate and an output plate. The code masking plate is arranged perpendicular to the axis of the electron gun and combinations of a binary code are laid out as punched holes. In order to simplify illustration, the code masking plate in Fig. 8 shows only 3 vertical and 8 horizontal rows. The vertical rows correspond to the 3 digits, which are necessary to represent $2^3 = 8$ discrete amplitude levels or states. The actual coding tube built by Bell Telephone Laboratories has a coding plate with 7 vertical and 128 horizontal rows, corresponding to 128 discrete amplitude levels.[10]

The voltage, obtained from the holding circuit, is applied to the Y deflector plates. When first applied, the beam is moved upward or downward in the left-hand unperforated region of the aperture plate into a position which corresponds to the applied amplitude. Then, by means of the quantization grid, the beam is shifted exactly into the horizontal row which is nearest to this position. After quantizing, the beam, now properly aligned, is swept in horizontal direction across the aperture plate. To perform this sweep, a linear sweep voltage, produced by a local synchronized sweep oscillator, is applied to the X deflector plates. Having completed the horizontal sweep, the beam is blanked and retraced to its starting position.

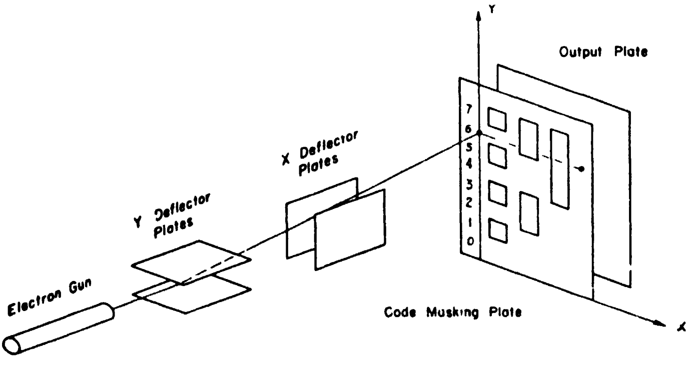


FIG. 8 Schematic of coding tube. (*Courtesy Bell Telephone Laboratories.*)

During the horizontal sweep, electrons which pass through the holes of the aperture plate are caught by the output plate and form a pattern of on-off pulses, Fig. 9, line A. This line shows the pattern obtained during 2 sweeps with a 3-digit aperture plate. The patterns represent the binary notations 010 and 011, which—reading backwards—correspond to the amplitude levels No. 2 and 6 (state numbers) of a scale with

8 discrete amplitude levels. The horizontal sweep is performed in such a way that the on-off pulses are equally spaced within the allotted time interval τ_0. With r digits, the time distance between adjacent pulses is thus reduced from $1/2B$ seconds to $1/2rB$ seconds. The coding tube therefore produces pulses at a rate of $2rB$ (24,000, $B = 4000$ cps, $r = 3$) pulses per second

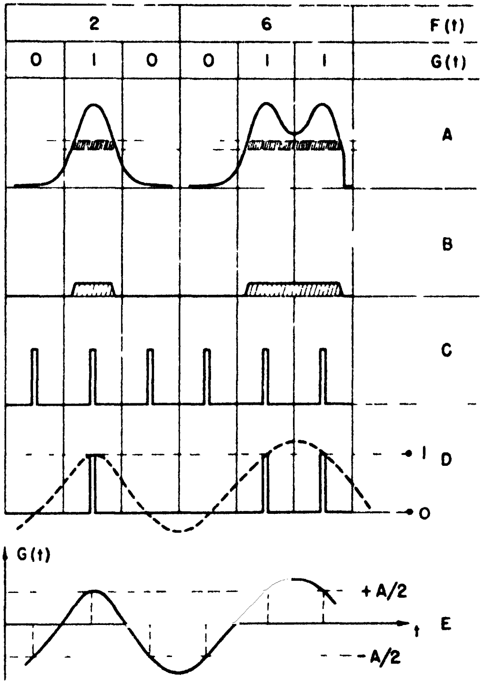


FIG. 9 Waveforms of pulses

These pulse patterns form the raw material for the secondary signal function $G(t)$ to be transmitted over the channel. Because of the increased pulse rate, the secondary signal $G(t)$ will have an increased signal bandwidth, rB instead of B. In other words, the number r of digits used with the aperture plate is the band expansion factor of the PCM system.

d. Quantizing. As the input pulses at the coding tube have a continuous amplitude range and because the beam has a finite cross section, it could happen that the beam straddles between 2 rows and thus sweeps out a combination of 2 adjacent codes. This is prevented by a quantization grid, Fig. 10. This grid consists of parallel wires aligned with the horizontal rows of the aperture plate. If the beam strikes a grid wire, secondary electrons are released and captured by the collector. The grid current is larger if the center of the beams strikes a wire and smaller if the beam is centered between 2 adjacent wires.

Figure 11, left, indicates that the grid current I_g consists of a D-C component and an A-C component, which varies periodically with the position y of the beam along the Y axis. The A-C component is amplified and applied in feedback relation to the amplifier, which produces the deflection voltage at the Y deflection plates.

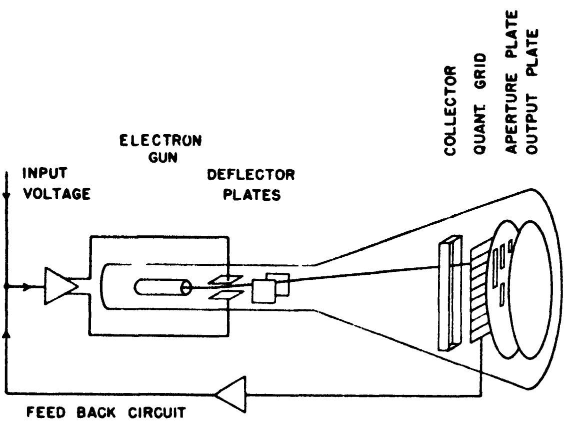


FIG. 10. Coding tube with quantization grid (*Courtesy Bell Telephone Laboratories.*)

The beam deflection voltage is equal to the sum of signal voltage v and feedback voltage $a \cdot \sin y$, and produces a deflection y proportional to this sum. Thus the resulting deflection y may be described by

$$y = v + a \cdot \sin y$$

where a is a measure for the degree of feedback, all unimportant constants of proportionality neglected.

Figure 11 shows the beam position y as a function of the input signal voltage v, with and without feedback. Without feedback ($a = 0$) the position is described by $y = v$. In other words, a signal voltage $v_1 = OA_1$ shifts the beam into position B_1, and a voltage $v_2 = OA_2$ shifts the beam into position B_2. With feedback, the position is determined by the curve $y = y(v)$, which has alternating positive and negative slopes. In general one obtains now more than one possible position for a given input voltage v. As an example, the voltage OA_1 or OA_2 can produce 3 different positions y. But positions on a negative slope are unstable and the beam jumps immediately into a stable position. As an example, the signal voltage $v_1 = OA_1$ corresponds to the point P_1 which is situated on a negative slope and which therefore jumps upward or downward into a stable point, say Q_1, which corresponds to the beam position C_1. Points exactly situated on the crossing points $P_1P_2 \cdots$ can jump upward or downward, but points to the left of $y = v$ can jump only upward, and points to the right can jump only in a downward direction.

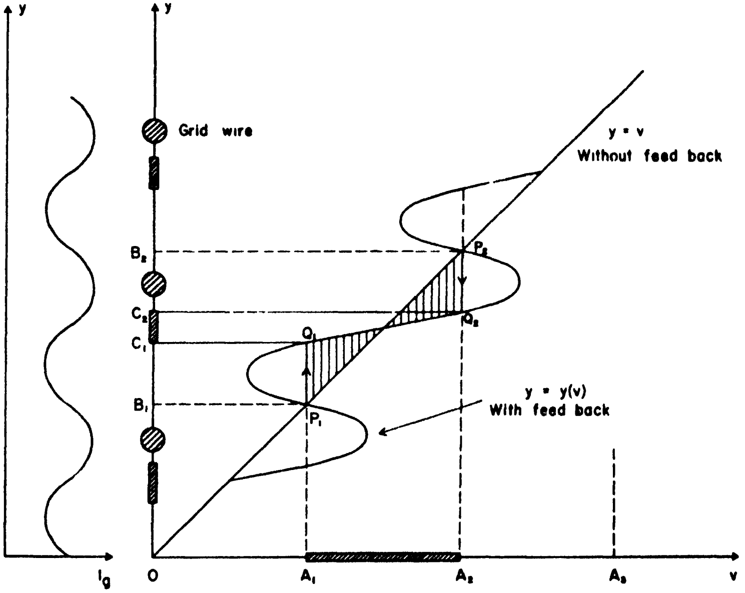


FIG. 11. Graphical representation of quantization.

As already mentioned, the signal voltage v is applied at first with the feedback circuit inoperative. Any voltage in the voltage range A_1A_2 will therefore correspond to a particular point on the straight line between P_1P_2, this means to a beam position between B_1B_2. When the feedback circuit is activated, the position curve is no longer given by straight section P_1P_2 but by the section Q_1Q_2. The beam therefore jumps upward or downward into a position between C_1C_2, and this range is much smaller than the original range B_1B_2. All amplitudes between OA_1 and OA_2 will therefore produce a beam position in a small range C_1C_2 just below the second grid wire. Signal voltages within OA_1 will likewise produce beam positions below the first grid wire and voltages within the range A_2A_3 will produce beam positions below the third grid wire, and so on.

The continuous signal range is thus transformed into a discontinuous range of fixed beam positions.

The quantization grid and the aperture plate are arranged in such a way, that the beam, if in its quantized position, illuminates fully the corresponding horizontal row of the aperture plate. As the beam is now swept across the aperture plate, it remains pressed against the lower surface of its guiding wire and cannot leave the once chosen row.

All signal amplitudes within a particular range are thus transformed into the same code group, as determined by the holes of the assigned horizontal row of the aperture plate.

e. Pulse Shaping. Figure 9, line *A*, shows 2 successive groups of code pulses, obtained with 2 horizontal sweeps in the case of $r = 3$ digits or $s = 8$ states. These pulses have now to be standardized in height, position, and duration. This is achieved by a circuit which performs two pulse-shaping operations, slicing and gating.

First, a thin slice is cut out from the train of pulses by a slicer circuit, Fig. 9, line *B*. The slicer circuit consists of a multivibrator or of a combination of two diodes. Bias conditions are so adjusted that the circuit is inoperative except when the signal falls through a narrow voltage range at about half the pulse amplitudes. These thin slices are amplified to a proper amplitude and are then gated by means of a pulse series, Fig. 9, line *C*. The gating pulses are produced by a local pulse generator with a pulse repetition frequency of $2rB$ cps (24,000 cps). After gating, the signal consists of a sequence of on-pulses (1) and off-pulses (0) in regular time intervals, $1/2rB$ seconds apart. The pulses are of equal height, either 0 or 1, of equal short duration and at fixed time positions.

The transmission of these sharp pulses would require a very broad frequency band. In order to save channel bandwidth, the pulses are used to produce the corresponding continuous signal of minimum bandwidth. According to the sampling theorem, the discrete amplitude levels, (0,1) in connection with the regular time interval $1/2rB$ seconds determine completely a continuous signal function with 2 states and a signal frequency band of rB cycles, the secondary signal $G(t)$. This signal is derived from the pulses in the way that $F(t)$ was derived from the sampling pulses in Fig. 2. That is, one sends the pulses through a low pass of bandwidth rB, Fig. 7. The result is a continuous signal—dotted line in Fig. 9, line *D*—which consists of the desired secondary signal and a d-c component. The signal is separated from the d-c component and amplified by the video-transmitting amplifier to a desired signal power S. As the signal $G(t)$, Fig. 9, line *E*, is limited to the bandwidth rB, it can be completely described by discrete ordinates $\pm A/2$, successive ordinates $1/2rB$ seconds apart. According to eq. 4, the signal power S in a unit resistance is determined by these discrete ordinates. As these ordinates are either $+A/2$ or $-A/2$, the signal power will be $S = A^2/4$.

$G(t)$ is the signal, which is sent into the channel instead of the original signal $F_0(t)$.

VII. Principal Operations at the Receiving End

a. Diagram. A simplified diagram of the receiving end in the case of a single PCM circuit is shown in Fig. 12. The transmitted signal $G(t)$ arrives at the receiver, attenuated by the transmission channel and accompanied by channel noise. A line filter keeps out all noise frequencies outside the signal band rB. As the accompanying noise contains now only frequencies in the band $0 \cdots rB$ cycles, it can—like the signal—be completely described by discrete noise ordinates, or noise samples, $1/2rB$ seconds apart, which coincide with the signal samples. Signal and noise are then amplified to a suitable power by the video receiving amplifier. For the sake of simplicity, one can assume that one finds at the output terminals of this amplifier the same signal power S as was delivered to the channel at the transmitting end. Together with the signal one will find a certain noise power rN, N being the noise power in the original signal band B.

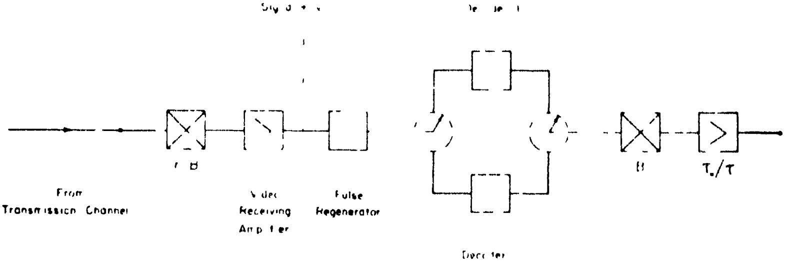


Fig. 12. Receiving end of a PCM circuit

b. Pulse Regeneration. Figure 13, line A, shows the same signal as Fig. 9, line E, but now disturbed by noise. The ordinates of the undisturbed signal (dotted line) are shifted in an upward or downward direction, depending on the instantaneous amplitude of the noise. The amplitudes at the sampling points are therefore no longer $\pm A/2$, but larger or smaller in a random manner.

By means of a slicing circuit, thin slices are cut out from the disturbed signal, Fig. 13, line B. These slices are then amplified and gated at the midpoints of their proper time intervals with narrow pulses, obtained by a local pulse generator, Fig. 13, line C. The result is a sequence of pulses, Fig. 13, line *D*, which is a true replica of the original sequence at the transmitter, Fig. 9, line *D*. This is true under the sole condition that the noise amplitudes *at the sampling points* do not exceed the signal amplitudes $\pm A/2$. If the noise amplitude *exceeds* the signal amplitude and is at the same time of *opposite* polarity, then one obtains a pulse instead of no pulse or no pulse instead of a pulse. These conditions tolerate a considerable amount of noise with no effect on the signal at all.

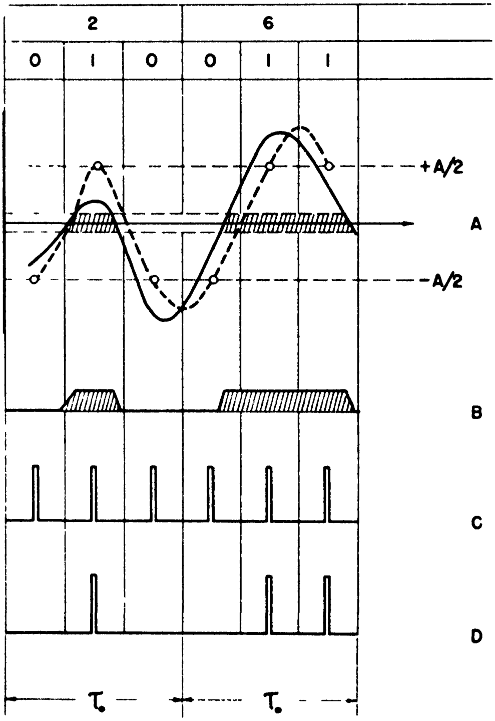


FIG. 13. Regeneration of pulses.

In the case of an intermediate repeater station, the pulses are again transformed into the continuous signal $G(t)$ and sent out over the second link, and so on. As pulse regeneration takes place at all intermediate stations, the pulses are at the final receiver as good as new. All channel noise will be cleaned out, independent of the number of links, provided the noise amplitude at any sampling point does not exceed the signal amplitude.

This series of signal pulses has now to be transformed into the primary signal $F(t)$.

c. Decoding. The purpose of the decoder is to retransform each group of pulses into one single pulse of corresponding amplitude, as an example the group 010 into amplitude 2 and the group 011 into amplitude 6, Figs. 9 and 13.

The decoding circuit, originally proposed by Shannon, consists mainly of a resistance-capacitor circuit *RC*, Fig. 14. By means of gate 1, which acts as a switch, the incoming pulses connect the *RC* circuit with a constant current source during the short duration of the pulses. In the simplest case, the constant current source consists of a battery and a very large resistor *W*, which provides for a constant charging current during the pulse duration, independent of the already existing charge at the condenser. Assume that the condenser has no charge when the pulse sequence, Fig. 15, begins. During the duration of the pulse in the second time interval the condenser obtains a fixed charge, and the voltage at the condenser will rise to 8 units (in general $s = 2^r$ units with an r-digit code). During the free time interval between successive pulses, the circuit *RC* is disconnected from the battery and the condenser discharges through resistor *R*. The time constant *RC* is such that the voltage drops

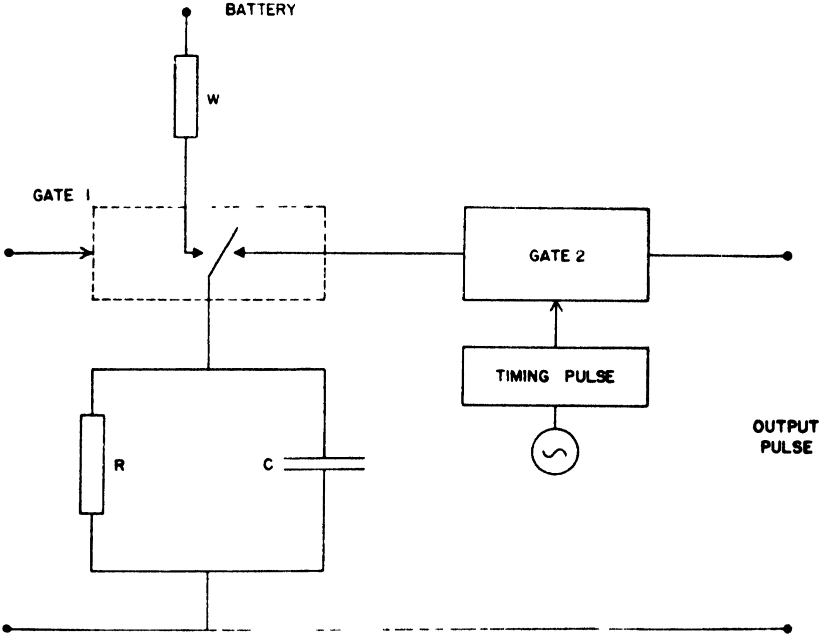


FIG. 14. Decoding circuit.

exactly by a factor of ½ in each time interval. Therefore, the voltage drops to 4 units at midpoint of the third time interval and to 2 units at midpoint of the fourth time interval. At this moment, gate 2 is opened for a very short time by a gating pulse and the result is one single pulse of amplitude 2 at the output terminals of the decoder. Thus the pulse group 010 = 2 delivers an output sample of 2 units.

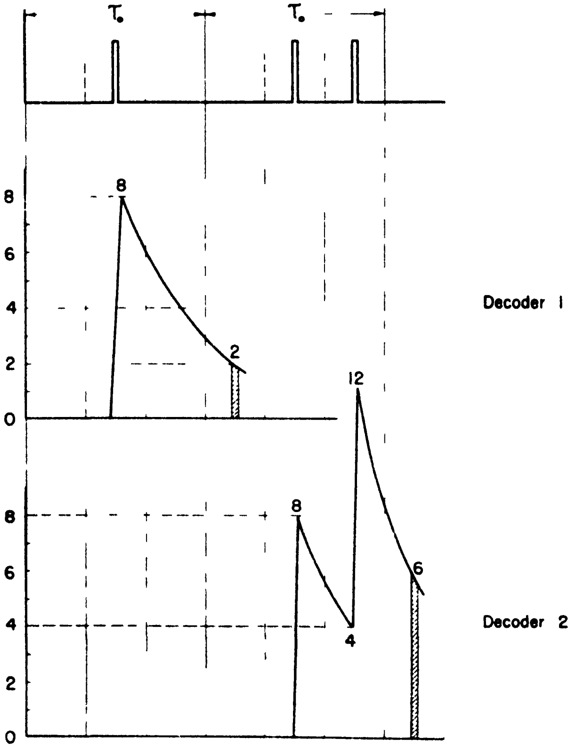


FIG. 15 Graphical representation of decoding

As the gating occurs in the first time interval of the second group, which may be occupied by a pulse, one has to use 2 decoders, one for the odd pulse groups and one for the even pulse groups, Fig. 12. These 2 decoders are connected with the incoming and outgoing lines by means of 2 rotating switches, which rotate with a frequency $1/2\tau_0 = B$ cps (4000 cps). One has thus sufficient time to discharge the condenser after the operation of the second gate, Fig. 14.

The decoder 2 decodes the second pulse group 011 = 6 in a similar way. The pulse in the second time interval charges the condenser to

8 units. The voltage has dropped to 4 units at the beginning of the pulse in the third interval and will be charged to $4 + 8 = 12$ units at the end of this pulse. This voltage drops to 6 units at midpoint of the next time interval. Here again, gate 2 of decoder 2 is opened and a pulse of amplitude 6 appears at the output terminals of the decoder. It may be seen that the decoder transforms each code group consisting of r pulses (binary number with r digits) into the corresponding absolute number, expressed by one pulse of corresponding amplitude.

At the output terminals of the decoding circuit, one obtains a series of pulses, spaced at regular time intervals $\tau_0 = 1/2B$ seconds, of equal short duration τ and of $s = 2^r$ discrete amplitude levels $0,1,2 \cdots (s - 1)$. Therefore, if one sends these pulses through a low pass of bandwidth B and amplifies with a linear gain τ_0/τ, one obtains, according to the sampling theorem, the continuous signal function $F(t)$ at the output terminals of the circuit.

As already mentioned, the recovered signal $F(t)$ differs somewhat from the original signal $F_0(t)$ which was produced by the source, owing to the quantization of $F_0(t)$. These random differences, which give rise to quantization noise, eq. 8, can be kept well within tolerable limits by a sufficiently large number of quantization steps at the encoding tube. In tolerating this small amount of noise, which is independent of the length of the circuit, one gets rid of the channel noise, as long as this noise remains below a comparatively large threshold value at which discrimination between mark and space becomes impossible. By pulse regeneration at the end of each link one not only cleans out the noise, but in doing so, one prevents even a small channel noise from accumulating over the many links of a long distance circuit to such an amount which eventually would make transmission of intelligence impossible at all.

VIII. FIDELITY IN PCM TRANSMISSION

With white noise, where no sharp threshold noise power exists, there is always a finite probability that noise exceeds the signal and thus may produce transmission faults. A transmission fault occurs if two things happen at the receiver at the same time: the noise amplitude has to exceed the signal amplitude at the sampling point and both amplitudes have to be of opposite polarity, Fig. 13, line A. If both are of equal polarity, no fault will occur, no matter how large the noise amplitude is. The error probability can therefore not exceed the value ½.

a. Transition Probabilities. The effect of channel noise on the signal can be described by transition probabilities, Fig. 16.

At the input of the encoder, the primary signal to be transmitted has s states, represented by s inputs $0,1 \cdots (s - 1)$. The source selects one state after another, and the transmission system intends to transmit the selected states to the far end of the circuit. The secondary signal, after encoding, has only two states, 0 and 1. A group of $r = \log s$ on-off pulses is sent out for each particular state of the primary signal. After transmission, these groups are decoded into the original states by the decoder. With ideal transmission, any particular state of the s possible states at the transmitter will produce the same state at the receiver, but none of the others This is only possible, if nothing happens to the on-off pulses during transmission over the channel.

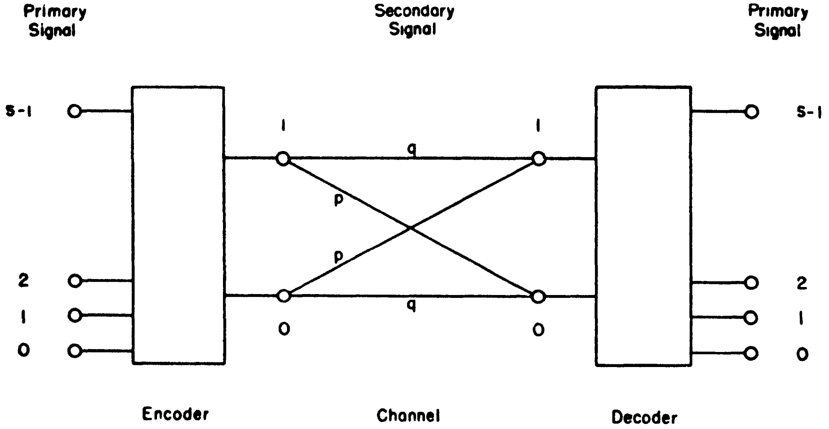


FIG. 16. Graphical representation of PCM transmission in the presence of noise.

TABLE III Transition probabilities

Transm. Sequence	011	011	011	011	011	011	011	011	= State 6
Received sequence	000	100	010	110	001	101	011	111	
Received state	0	1	2	3	4	5	6	7	
Probability	qpp	ppp	pqp	pqp	qpq	ppq	qqq	pqq	

Due to the channel noise, there is a certain probability p that one sends out state 1 but receives state 0 and vice versa. Consequently, the probability that no transmission fault occurs is q, and $p + q = 1$. With no noise, one has $p = 0$ and $q = 1$. With infinite noise one has $p = q = \frac{1}{2}$. This is also intuitively the worst case, as the chance of obtaining 1 (or 0) is the same whether one sends 1 or 0.

The question is now, how noise will effect transmission of the primary signal between the input terminals of the encoder and the output terminals of the decoder, if the transition probabilities p and q with the secondary signal are known. The secondary signal to be decoded at the receiver consists of a series of on-off pulses. The fraction p of the pulses is wrong and the fraction q is right. The faults are distributed in a random manner within the sequence. As now groups of such pulses are decoded into the primary samples, such a primary sample will be wrong whenever a fault occurs in the pulse group. If one transmits with a 3-digit code for example state $6 = 011$, the probability of receiving $110 = 3$ will be $pqp = p^2q$.

Table III shows the probabilities that any one of the 8 possible states will be received with a 3-digit code if nothing but a sequence of state 6 is sent. Again, without noise, $p = 0$ and $q = 1$, one receives only state 6 and none of the others. With an infinite noise, $p = q = \frac{1}{2}$, all probabilities are equal ($\frac{1}{8}$). That is, one sends state 6 in sequence, but receives a random sequence of all possible states, without any relation to the transmitted state. Such a random sequence would be considered as a noise, with no content of signal whatsoever.

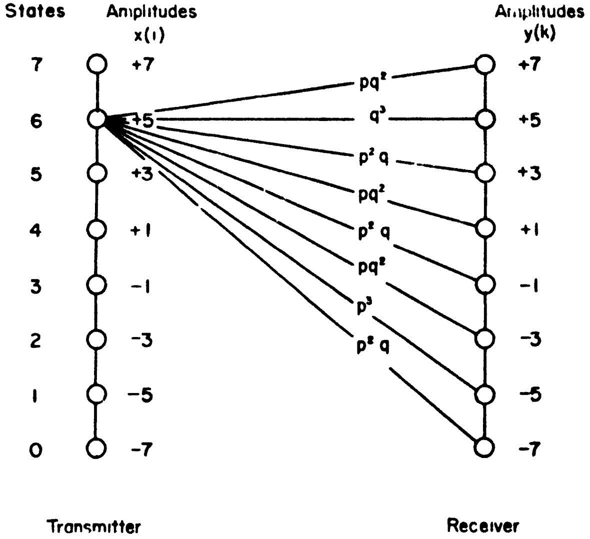


FIG. 17 Set of transition probabilities

Figure 17 shows the considered set of transition probabilities emerging from state 6. Similar sets emerge from all other states. Each state corresponds to a discrete signal amplitude $x(i)$ at the receiver, measured in arbitrary units, and correspondingly one has the same set of discrete amplitudes $y(k)$ at the receiver. No others than these discrete amplitudes $y(k)$ can be produced by the decoder, no matter what happens on the channel. The transition probability $p_i(k)$ is the probability that one receives the amplitude $y(k)$ if one sends the signal amplitude $x(i)$.

The following general relations hold:

$$\sum_i p_i(k) = \sum_k p_i(k) = 1 \tag{13}$$

$$\sum_k p_i(k) \cdot y(k) = (q - p) \cdot x(i) \tag{14}$$

Equation 14 states that the center of gravity of the weighted received amplitudes is proportional to the transmitted signal amplitude. The center of gravity is still a measure for the transmitted signal, even in the presence of channel noise.

If one sends $x(i)$ and receives $y(k)$, one can assume that one received the *signal* amplitude $x(i)$ together with a *noise* amplitude $y(k) - x(i)$. One can as well assume that the received signal amplitude was $m \cdot x(i)$ and the noise amplitude was $y(k) - m \cdot x(i)$, where m is a constant, independent of amplitude, which has to be determined in such a way that the noise power at the output becomes a minimum.

A measure of fidelity is the mean square of the Euclidean distance between the fixed output signal amplitude $m \cdot x(i)$ and the possible output amplitudes $y(k)$, taking into account their probabilities, or

$$d_i^2 = \sum_k p_i(k) \cdot [y(k) - m \cdot x(i)]^2 \tag{15}$$

Obviously, d_i^2 is the noise power at the output terminals of the decoder in case one sends the signal amplitude $x(i)$ into the encoder.

This noise power has to be minimized, $(\partial/\partial m)(d_i^2) = 0$. This determines the factor m and one finds by means of equation 14

$$m = q - p \tag{16}$$

That is, in order to obtain minimum noise power—or maximum signal power—one has to measure the Euclidean distances against the center of gravity.

Equation 15 may now be written

$$\sum_k p_i(k) \cdot y(k)^2 = (q - p)^2 \cdot x(i)^2 + d_i^2 \tag{17}$$

The left-hand side of eq. 17 is the output power if one sends a fixed signal power $x(i)^2$. The right-hand side indicates that the output power consists of two parts, the received signal power $(q - p)^2 \cdot x(i)^2$ and the noise power d_i^2.

Equation 17 shows clearly the effect of channel noise on the transmitted signal. Any possible signal amplitude $x(i)$ at the transmitter will be found—in a statistical sense—with a reduced amplitude $(q - p)$.

$x(i)$ at the receiver. In the course of time all possible signal amplitudes will occur at the input terminals of the encoder. If $p(i)$ is the probability of amplitude $x(i)$, the signal power at the transmitting end will be

$$S = \sum_i p(i) \cdot x(i)^2 \tag{18}$$

The total power at the receiving end becomes, eq. 17,

$$P = \sum_i \sum_k p(i) \cdot p_i(k) \cdot y(k)^2 \tag{19}$$

This power is partly signal power and partly noise power. The received signal power is

$$S^* = (q - p)^2 \cdot \sum_i p(i) \cdot x(i)^2 = (q - p)^2 \cdot S \tag{20}$$

and the received noise power will be

$$N^* = \sum_i p(i) \cdot d_i^2 = P - S^* \tag{21}$$

Equation 19 shows that the total output power P depends on the statistical structure of the primary signal and on the transition probabilities. In general, the power P is different from the power S at the transmitter, eq 18. But it is easy to see that P approaches S in case the error probability approaches zero. Furthermore, if the statistical structure of the signal at the transmitter is such that all possible amplitudes $x(i)$ occur with equal probabilities, then all possible amplitudes $y(k)$ at the receiver occur also with equal probabilities, or $P = S$.

In these important cases, one obtains the simple equations

$$S^* = (q - p)^2 \cdot S = (1 - 4pq) \cdot S \tag{22}$$

$$N^* = 4pq \cdot S \tag{23}$$

The output signal-noise ratio is therefore

$$\frac{S^*}{N^*} = \frac{1}{4pq} - 1 \tag{24}$$

The output signal-noise ratio drops from infinity with a noiseless channel ($p = 0$, $q = 1$) to zero in the case of an infinitely large channel noise ($p = q = \frac{1}{2}$).

So far, transmission over one link only was considered. If the PCM circuit consists of a certain number of links in tandem, transmission faults occur within each link and cumulate over the links. The probability p_n that a pulse will be wrong after transmission over n links increases therefore with the number n of links.

In the case of 2 links, for example, some pulses will get wrong on the second link which were right on the first link, but also some pulses which were wrong on the first link will become right again on the second link. Consequently, the probability that a pulse is received right after transmission over 2 links is $q_2 = q^2 + p^2$ and the probability that it will be wrong is $p_2 = pq + qp$. Therefore

$$q_2 - p_2 = (q - p)^2$$

or in general with transmission over n links

$$q_n - p_n = (q - p)^n$$

The received signal power after transmission over n channels will therefore be, eq 22,

$$S^* = (q - p)^{2n}S$$

and the accompanying noise power will be

$$N^* = S - S^* = [1 - (q - p)^{2n}] \cdot S$$

The resulting output signal-noise ratio is therefore

$$\frac{S^*}{N^*} = \frac{(q - p)^{2n}}{1 - (q - p)^{2n}} \tag{25}$$

b. Error Probability and Channel Noise. It remains now to establish a relation between error probabilities and channel noise, more exactly between error probability and the signal-noise ratio on the channel. This makes it then possible to express the output signal-noise ratio, eqs. 24, 25, in terms of the channel signal-noise ratio, and to see what noise-cleaning effect was obtained by PCM transmission.

With the notations of Section II, the channel noise power in the band B of the primary signal was N. With PCM transmission and an r-digit code, the channel bandwidth has to be expanded to rB and the channel noise will increase to rN. The channel signal-noise ratio is therefore $S:rN$, referred to one link.

As noise is restricted to the channel bandwidth rB, it can be described by discrete noise samples, $1/2rB$ seconds apart, which coincide with the signal samples. In order to establish the desired relation, one has not only to know the channel noise power, but also the statistical structure of the noise. Various types of noise, with equal noise power, but different statistical structure, give rise to different error probabilities.

In the important case of white noise, the statistical structure is given by a Gaussian error curve, Fig. 18. The probability of finding a noise sample of amplitude between V and $V + dV$ is

$$W(V)dV = \frac{1}{\sqrt{2\pi}\, V_0} \cdot e^{-V^2/2V_0^2} dV$$

V_0 is determined by the mean noise power, $V_0^2 = rN$.

Figure 18 shows also the 2 discrete signal amplitude levels $+A/2$ and $-A/2$. The signal power is $S = A^2/4$.

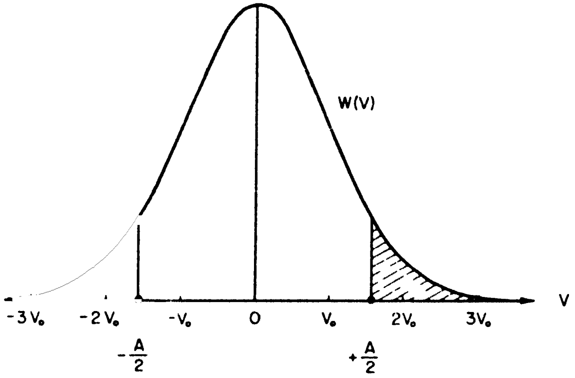


FIG 18 Sample distribution of white noise.

With any symmetrical distribution curve, the error probability is clearly

$$p = \int_{A/2}^{\infty} W(V)dV \tag{26}$$

In the case of white noise one obtains

$$p = \tfrac{1}{2}[1 - \phi(\sqrt{S/rN})]$$

and

$$q - p = \phi(\sqrt{S/rN}) \tag{27}$$

where ϕ is the well-known probability integral

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \cdot \int_{-x}^{+x} e^{-u^2/2} \cdot du$$

Equation 27 establishes the desired relation between error probability p and the signal-noise ratio S/rN on the channel. With this relation one obtains, eqs. 24, 25,

$$\frac{S^*}{N^*} = \frac{\phi^{2n}}{1 - \phi^{2n}} \tag{28}$$

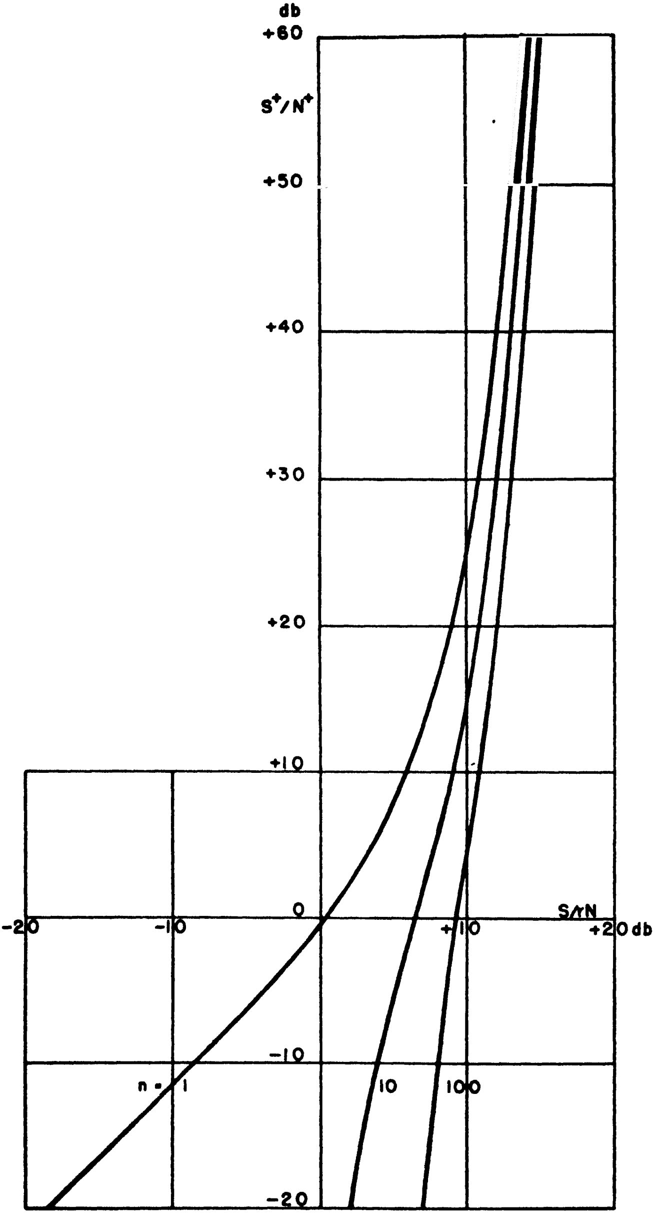


Fig. 19. Output signal-noise ratio.

Figure 19 shows, according to eq. 28, what output signal-noise ratio S^*/N^* will be obtained with a given channel signal-noise ratio S/rN. The 3 curves correspond to $n = 1$, 10, and 100 links in tandem. In the case of one link, no improvement will be obtained in case the channel noise power is about equal or larger than the signal power. But as the channel noise power becomes smaller, the output signal-noise ratio improves rapidly

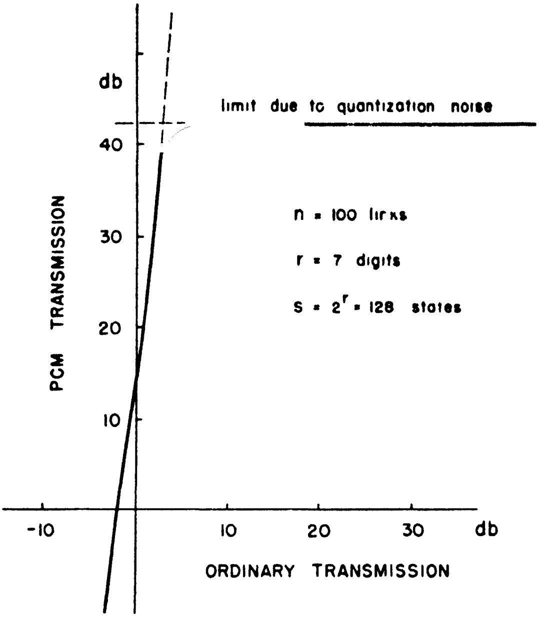


FIG. 20 Output signal-noise ratio with and without PCM.

In case the channel signal-noise ratio is already high, $S/rN \gg 1$, the approximation

$$\phi(x) \approx 1 - \sqrt{\frac{2}{\pi}} \cdot \frac{e^{-x^2/2}}{x}$$

may be used. In this case one obtains, $n = 1$

$$\frac{S^*}{N^*} \approx \sqrt{\frac{\pi}{8}} \sqrt{\frac{S}{rN}} \; e^{S/2rN}$$

S^*/N^* improves mainly exponentially with the signal-to-noise ratio S/rN on the channel.[15]

In the case of 10 links and even more in the case of 100 links, the output signal-noise ratio improves almost abruptly from very low to very high values. With 100 links, for example, a small increase of the channel signal-noise ratio from 10 db to 14 db improves the ratio at the output from about 4 db to 40 db. Furthermore, the requirements of the channel signal-noise ratio per link are almost independent of the number of links. 13 db in the case of one link and 15 db in the case of 100 links for 50 db output signal-noise ratio. This increase of 2 db accounts for the whole cumulating effect over 100 links.

With normal transmission, the output signal-noise ratio would be S/N with one link and S/nN with n links, as here noise accumulates with the number of links. Figure 20 compares this case with PCM, $r = 7$ digits, $s = 128$ states, $n = 100$ links. In both cases, the same signal power is used for transmission, and the channels have equal noise power per cycle. For example, if the noise conditions are such that one obtains with ordinary transmission a signal-noise ratio of 3 db, PCM will improve it to 40 db. Quantization noise limits the maximum obtainable signal-noise ratio in this case to 42 db, eq. 8.

One sees clearly that information can be transmitted over a PCM system under noise conditions, where ordinary transmission would completely fail. No other known transmission method can achieve such striking results.

IX. Rate of Transmission

Besides a loss in fidelity, information is also lost during transmission over a noisy channel.

According to eq. 10 one can feed the channel of a binary PCM system at the maximum with 2 bits per second and cycle band. The question is how many bits are lost as a result of the noise.

Assume again, as indicated in Fig 16, that one has $s = 2^r$ states with the signal to be transmitted before encoding. The encoder, channel, and decoder establish a connection between the input and output states. Without channel noise a particular input state is connected only with the corresponding output state, but with none of the others. With channel noise, any input state is connected with all output states by a set of transition probabilities $p_i(k)$. This is, if one sends a particular state (i) into the encoder, there is a finite probability $p_i(k)$ that one receives any other state (k) at the output terminals of the decoder. Noise makes it uncertain which state will actually be received.

According to Shannon,[17] a measure of the uncertainty related with the transmission of a particular state (i) is the entropy of the set of transition probabilities which emerges from this state, or

$$H_i = -\sum p_i(k) \cdot \log p_i(k)$$

It is easy to see that the entropy of the set in Fig. 17 is $-3(p \log p + q \log q)$. In this case one has $8 = 2^3$ states. With $s = 2^r$ states one finds the general relation

$$H_i = -r(p \cdot \log p + q \cdot \log q)$$

In other words, H_i does not depend on the particular state (i) at all. All states get the same treatment by the channel noise.

The average entropy, or equivocation due to noise is therefore

$$H(n) = \sum_i p(i) \cdot H_i = -r(p \cdot \log p + q \cdot \log q) \tag{29}$$

The equivocation has now to be compared with the entropy of the signal after decoding. If $p(k)$ is the probability of the occurrence of the output state (k), this entropy is

$$H(y) = -\sum_k p(k) \cdot \log p(k)$$

$H(y)$ has to be maximized under condition that the number s of possible states is fixed. A maximum of $H(y)$ is then obtained if all states occur with equal probabilities $p(k) = 1/s$, or

$$H(y) = \sum_1^s \frac{1}{s} \cdot \log s = r \tag{30}$$

It may be remarked that the maximizing condition $p(k) = 1/s$ is identical with the condition that all input states occur with equal probability $p(i) = 1/s$. Obviously the relation holds

$$p(k) = \sum_i p(i) \cdot p_i(k)$$

If all values of $p(i)$ are equal, $p(i) = 1/s$, one has simply, eq. 13,

$$p(k) = p(i) = 1/s$$

and also the signal entropies at the input and output are equal

$$H(x) = H(y) = r$$

There are now two sources at work, the information source which produces $H(y)$ bits of information per sample at the output terminals of the decoder and the noise source which works against the information source and which gives rise to the equivocation $H(n)$ bits per sample.

The difference between $H(y)$ and $H(n)$ is the rate at which information is transmitted, or

$$R = r(1 + p \cdot \log p + q \cdot \log q) \qquad (31)$$

All these equations are based on a per sample basis, and it may be remembered that any sample of the primary signal function carries exactly $r = \log s$ bits of information.

If B is the signal frequency band and consequently rB is the channel frequency band, one can send at a maximum $2B$ samples per second into the encoder. The rate of transmission on a per-second basis is therefore

$$R = 2rB(1 + p \log p + q \log q) \text{ bits per second}$$

Or the rate per second and cycle channel band will be

$$\frac{R}{r \cdot B} = 2(1 + p \cdot \log p + q \cdot \log q) \qquad (32)$$

Without noise, $p = 0$ and $q = 1$, one transmits 2 bits per second and cycle, which is the characteristic maximum for binary PCM. With infinite noise, $p = q = \frac{1}{2}$, one obtains $R = 0$ and no information is transmitted at all.

Due to the identities

$$q = \tfrac{1}{2}(1 + (q - p))$$
$$p = \tfrac{1}{2}(1 - (q - p))$$

one sees that the rate depends only on the difference $q - p$. This is also the case for the output signal-noise ratio $S^*:N^*$, eqs. 24 and 25. There exists therefore a unique relation between rate of information and fidelity, measured by the output signal-noise ratio.

This relation is shown in Fig. 21 and holds for any random cause of transmission errors With high signal-noise ratios the rate approaches the limit of 2 bits per second and cycle bandwidth, but drops with decreasing signal-noise ratio and approaches zero as the noise becomes infinite large.

In eq. 32 the value of p and q depend primarily on the channel signal-noise ratio $S:rN$. In the important case of white noise, p and q are determined by eq. 27.

Figure 22 shows the rate of transmission, now as a function of the channel signal-noise ratio. Curve C is the channel capacity, calculated by eq. 12. This is the maximum rate which can only be obtained by ideal encoding. The 3 curves designated with 1, 10, and 100 refer to PCM circuits consisting of 1, 10, and 100 links in tandem. One sees that with 13-15 db signal-noise ratio per link one is very close to the maximum rate of 2 bits in all 3 cases. This again indicates that the requirements on the channel signal-noise ratio per link are almost independent of the number of links, 13 db in the case of one link and 15 db in the case of 100 links. In the case of 10 links and even more in the case of 100 links, the rate drops almost abruptly to very small values. This behavior corresponds quite to the sharp drop in output-signal: noise ratio, Fig. 19.

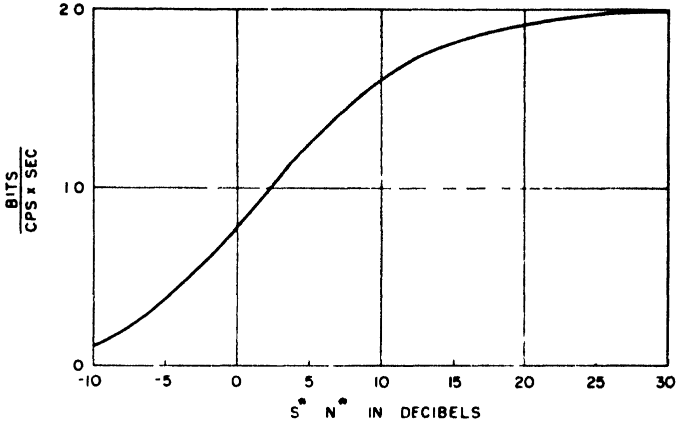


FIG. 21 Rate of information and output signal-noise ratio

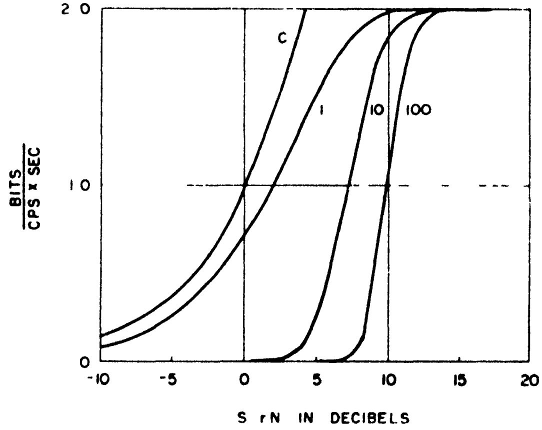


FIG 22 Rate of transmission and channel signal-noise ratio.

In order to transmit 2 bits per cycle and second one needs with ideal encoding a signal-noise ratio of 3:1 (5 db) on the channel, curve *C* in Fig. 22. With PCM and white noise one needs a signal-noise ratio of 20:1 (13 db). This means that with a more involved method of encoding one could still save a factor of about 7 in signal power. PCM, although it is more effective than any other known noise-cleaning method, still wastes 85% of the available signal power.

Unfortunately, as soon as one attempts to approach the ideal case, the transmitter and the receiver become extremely complicated. At the present time PCM seems to be a good economic balance between noise-cleaning efficiency and complexity of the encoding and decoding devices.

REFERENCES

1. Reeves, H. A. U.S. Patent 2,272,070, Feb. 3, 1942, assigned to International Standard Electric Corp.; also French Patent 852,183, Oct. 23, 1939.
2. Goodall, W. M. Telephony by pulse code modulation. *Bell System Tech. J.*, **26,** 395–409 (1947).
3. Black, H. S. Pulse code modulation. *Bell Labs. Record*, **25,** 265–269 (1947).
4. Grieg, D. D. Pulse count modulation. *Elec. Commun.*, **24,** 287–296 (1947).
5. Clavier, A. G., Panter, P. F., and Grieg, D. D. PCM distortion analysis. *Elec. Eng.*, **66,** 1110–1122 (1947).
6. Black, H. S., and Edson, J. O. PCM equipment. *Elec. Eng.*, **66,** 1123–1125 (1947); *Trans. Am. Inst. Elec. Engrs.*, **66,** 47–131 (1947).
7. *Electronics*, Coded pulse modulation minimizes noise. **20,** 126–131 (1947).
8. *Electronics*, Bandwidth *vs* noise in communication systems. **21,** 72–75 (1948).
9. Meacham, L. A., and Peterson, E. An experimental multichannel pulse code modulation system of toll quality. *Bell System Tech. J.*, **27,** 1–43 (1948).
10. Sears, R. W. Electronic beam deflection tube for pulse code modulation. *Bell System Tech. J.*, **27,** 44–57 (1948).
11. Bennet, W. R. Spectra of quantized signals. *Bell System Tech. J.*, **27,** 446–471 (1948).
12. Feldman, C. B. A 96 channel pulse code modulation system. *Bell Labs. Record*, **26,** 364–369 (1948).
13. Oliver, B. M., Pierce, I. R., and Shannon, C. E. The philosophy of PCM. *Proc. Inst. Radio Engrs.*, **36,** 1324–1331 (1948).
14. Shannon, C. E. Communication in the presence of noise. *Proc. Inst. Radio Engrs.*, **37,** 10–21 (1949).
15. Clavier, A. G., Panter, P. F., and Dite, W. Signal to noise ratio improvement in a PCM system. *Proc. Inst. Radio Engrs.*, **37,** 355–359 (1949).
16. Deloraine, E. M. Pulse modulation. *Proc. Inst. Radio Engrs.*, **37,** 702–705 (1949).
17. Shannon, C. E. A mathematical theory of communication, *Bell System Tech. J.*, July and October, 1949.

A Summary of Modern Methods of Network Synthesis

E. A. GUILLEMIN

Massachusetts Institute of Technology, Cambridge, Massachusetts

CONTENTS

Page
Introduction 261
I Analytic Form of an Impedance (Resp Admittance) and Its Real Part 262
II Conditions and Tests for Positive Real Character 263
III Some Important Properties of Hurwitz Polynomials and Positive Real Functions 264
IV Special Forms of $Z(\lambda)$ in the Two-Element Cases 265
  1 *LC* Networks (Lossless Case) 265
  2 *RC* or *RL* Networks 266
V. Some Remarks Relevant to the Brune Process 267
VI The Darlington Procedure for the Solution of the Brune Problem Skeletonized 271
VII. Synthesis of the Single-Loaded Lossless Coupling Network for a Prescribed Magnitude of Transfer Impedance 275
VIII Cauer's Method of Synthesis from a Specified $|Z_1(j\omega)|^2$ 276
IX. Complementary Impedances, Constant-Resistance Filter Groups 279
X Another Way of Designing for Finite Resistances at Both Source and Load 281
XI The Constant-Resistance Lattice 283
XII An Alternate Realization Procedure for Transfer Functions 286
XIII Synthesis of a Lossless Two Terminal-Pair Network through the Ladder Development of z_{22} 286

ILLUSTRATIVE EXAMPLES

XIV. Brune's Synthesis Procedure 290
XV. Darlington's Procedure Applied to the Same Problem 292
XVI An Alternative Method of Synthesis that Avoids Mutual Coupling 293
XVII. Darlington's Procedure Applied to the Synthesis of a Transfer Impedance 295
XVIII. Cauer's Method Applied to the Same Problem 296
XIX A Constant-Resistance Filter Group 297
XX The Same Transfer Function Realized through a Lossless Network with Resistance Loading at Both Ends 298
XXI Realization through a Cascade of Amplifier Stages 299
XXII Further Illustration of the Ladder Development Procedure 300
References 303

INTRODUCTION

The theory of lumped-constant linear passive network synthesis is of too recent origin to have reached a stage of adequate documentation. As a result there exists, among those interested in its application to practical problems, considerable inaccurate understanding of the basic principles underlying this theory. The following very compact summary of the most essential principles and procedures, together with some illustrative examples, may help to clarify this situation. It is assumed that the reader has a reasonably good general background and is more in need of concise statements than detailed elaboration and orientation.

As regards selection of material, preference to some extent is given those subjects for which the treatment in the existing literature is less satisfactory. The following discussion is in no way intended as a complete or exhaustive treatment, but represents, in connected sequence, a selection of material forming the main stem of the present-day synthesis theory as it applies to passive lumped-constant networks.

I. ANALYTIC FORM OF AN IMPEDANCE (RESP. ADMITTANCE) AND ITS REAL PART

A rational function of the complex frequency variable $\lambda = \sigma + j\omega$ is the driving point impedance*

$$Z(\lambda) = \frac{P(\lambda)}{Q(\lambda)} = \frac{a_0 + a_1\lambda + a_2\lambda^2 + \cdots + a_n\lambda^n}{b_0 + b_1\lambda + b_2\lambda^2 + \cdots + b_n\lambda^n} = \frac{m_1 + n_1}{m_2 + n_2} \tag{1}$$

For $\lambda = j\omega$, $P(-\lambda)$ is the conjugate of $P(\lambda)$ and $Q(-\lambda)$ is the conjugate of $Q(\lambda)$. Hence the familiar process of rationalization applied to the denominator of $Z(\lambda)$ is equivalent to multiplying numerator and denominator by $Q(-\lambda)$, thus

$$Z(\lambda) = \frac{P(\lambda)\cdot Q(-\lambda)}{Q(\lambda)\cdot Q(-\lambda)} = \frac{(m_1 + n_1)(m_2 - n_2)}{(m_2 + n_2)(m_2 - n_2)} \tag{2}$$

One finds

$$\mathrm{Re}[Z(j\omega)] = \left[\frac{m_1m_2 - n_1n_2}{m_2^2 - n_2^2}\right]_{\lambda = j\omega} = \frac{A_0 + A_1\omega^2 + A_2\omega^4 + \cdots + A_n\omega^{2n}}{B_0 + B_1\omega^2 + B_2\omega^4 + \cdots + B_n\omega^{2n}} = \frac{A(\omega^2)}{B(\omega^2)} \tag{3}$$

Here

$$\begin{aligned} A(-\lambda^2) &= m_1m_2 - n_1n_2 = \text{even part of } P(\lambda)\cdot Q(-\lambda) \\ B(-\lambda^2) &= m_2^2 - n_2^2 = Q(\lambda)\cdot Q(-\lambda) \end{aligned} \tag{4}$$

In factored form

$$\begin{aligned} A(-\lambda^2) &= A_n(\lambda_1^2 - \lambda^2)(\lambda_2^2 - \lambda^2)\cdots(\lambda_n^2 - \lambda^2) \\ A(\omega^2) &= A_n(\lambda_1^2 + \omega^2)(\lambda_2^2 + \omega^2)\cdots(\lambda_n^2 + \omega^2) \end{aligned} \tag{5}$$

---

* Here m_1 and n_1 are respectively the even and odd parts of $P(\lambda)$, while m_2 and n_2 are the even and odd parts of $Q(\lambda)$.

Note that $A(-\lambda^2)$ is a polynomial in λ^2 since it contains only even powers of λ. The factors in eqs. 5 place the λ^2-roots in evidence. The reason for writing these factors as $(\lambda_\nu^2 - \lambda^2)$ instead of $(\lambda^2 - \lambda_\nu^2)$ will be seen in the next article.

II. CONDITIONS AND TESTS FOR POSITIVE REAL CHARACTER*

Concise Statement: The necessary and sufficient conditions that a rational function $Z(\lambda)$, which is real for real λ, be a positive real (p.r.) function are: (a) $Z(\lambda)$ must be analytic in the right-half λ-plane; (b) $\text{Re}[Z(j\omega)] \geq 0$ for all real values of ω. In case $Z(\lambda)$ has poles for $\lambda = j\omega$ one must add: (c) Any j-axis poles must be simple, and the residues of $Z(\lambda)$ must there be real and positive These are the "ABC's" of a positive real function.

The preliminary requirement that $Z(\lambda)$ be real for real λ means simply that the polynomials $P(\lambda)$ and $Q(\lambda)$ must have *real* coefficients.

Regarding the process of testing $Z(\lambda)$ to see if the ABC's are met, we note that (a) requires $Q(\lambda)$ to be a Hurwitz polynomial (abbreviated H.P.). This test upon $Q(\lambda)$ will reveal any j-axis poles of $Z(\lambda)$ and hence will show whether or not (c) is relevant to the specific case at hand.

Relative to the test for (b) we note from eqs. 3 and 5 that $A(\omega^2)$ and hence $\text{Re}[Z(j\omega)]$ is positive for all ω if $A(-\lambda^2)$ has no negative real λ^2-roots of odd multiplicity, or if $A(\omega^2)$ has no positive real ω^2-roots of odd multiplicity. Sturm's theorem is the method of testing for this possibility. (Incidentally, the coefficient A_n in eqs. 3 or 5 must obviously be positive to insure $A(\omega^2)$ positive for $\omega^2 \to \infty$; and the convenience that results from writing the factors in eqs. 5 as they are, should now be evident since one would otherwise have to introduce a factor $(-1)^n$ in this expression.) This test, which assures that $A(\omega^2)$ has only even multiplicity zeros for real ω's (if it has any such zeros at all) amounts to making sure that if the real part of $Z(\lambda)$ becomes zero somewhere on the j-axis, this zero shall be a *minimum* of the real part.

In the event that $Z(\lambda)$ has any j-axis poles, these are revealed as common factors of m_2 and n_2; they are factors of the form $(\lambda^2 + \omega_\nu^2)$. The residue of $Z(\lambda)$ at such a pole may be found from

$$k_\nu = \left(\frac{m_1 + n_1}{m_2' + n_2'}\right)_{\lambda = j\omega_\nu} \tag{6}$$

in which the primes on the m_2 and n_2 denote differentiation with respect to λ. Now if we look at the expression

$$A(-\lambda^2) = m_1 m_2 - n_1 n_2 \tag{7}$$

---

* In this connection see also Arts. 26 and 27 of Chapter VI in The Mathematics of Circuit Analysis by E. A. Guillemin, John Wiley and Sons, 1949.

we see that the point $\lambda = j\omega_\nu$ is a j-axis zero, and the test for (b) has required that such zeros in $A(-\lambda^2)$ be of *even* multiplicity. Thus the factor $(\lambda^2 + \omega_\nu^2)$ must be contained at least *twice* in $A(-\lambda^2)$, so that not only this function but also its first derivative is zero at $\lambda = j\omega_\nu$, that is,

$$m_1m_2' - n_1n_2' + m_1'm_2 - n_1'n_2 = 0 \text{ for } \lambda = j\omega_\nu \tag{8}$$

Since the last two terms are separately zero, we have

$$(m_1m_2' - n_1n_2')_{\lambda=j\omega_\nu} = 0 \tag{9}$$

or

$$\frac{m_1}{n_2'} = \frac{n_1}{m_2'}, \text{ for } \lambda = j\omega_\nu \tag{10}$$

and so the residue as given by eq. 6 becomes*

$$k_\nu = \left(\frac{m_1}{n_2'}\right)_{\lambda=j\omega_\nu} = \left(\frac{n_1}{m_2'}\right)_{\lambda=j\omega_\nu} \tag{11}$$

Noting that n_2' is even and m_2' is odd, we see that k_ν is *real* in any case. This realness of k_ν is assured by the requirement (b). In other words, if the real part of $Z(\lambda)$ is positive for $\lambda = j\omega$ then at any simple j-axis poles, $Z(\lambda)$ is assured to have real residues, but we must still compute them according to eq. 11 to see if they are positive.

Since the Hurwitz test on $Q(\lambda)$ and the Sturm test of $A(\omega^2)$ plus evaluation of residues in the event of j-axis poles is in general a tedious process, one should first apply to $Z(\lambda)$ some simple tests (even though insufficient) to weed out quickly any functions that obviously cannot be p.r. Such requirements on $Z(\lambda)$, which can be tested by inspection, are: (a) $P(\lambda)$ and $Q(\lambda)$ can have no negative coefficients; and they can have no missing terms except in the special case that they degenerate into even or odd functions. (b) The highest and the lowest powers of $P(\lambda)$ and $Q(\lambda)$ can differ at most by unity. All that has been said applies equally well to an admittance; and if a given rational function is p.r., its reciprocal is p.r. also.

III. SOME IMPORTANT PROPERTIES OF HURWITZ POLYNOMIALS AND POSITIVE REAL FUNCTIONS

If $h(\lambda) = m(\lambda) + n(\lambda)$ is a Hurwitz polynomial then $m(\lambda)$ and $n(\lambda)$ have simple zeros alternating on the j-axis. The rational function m/n or n/m has simple poles restricted to the j-axis and positive real residues.

---

* It may happen that the polynomial in eq. 9 contains the factor $(\lambda^2 + \omega_\nu^2)$, in which case one of the expressions in eq. 11 becomes indeterminate and invalid.

If $Z(\lambda) = P(\lambda)/Q(\lambda)$ is a p.r. function then $P(\lambda)$ is revealed to be a Hurwitz polynomial. Note that the ABC's require only the Hurwitz test of $Q(\lambda)$, but once the p.r. character of $Z(\lambda)$ is established, $P(\lambda)$ is proved also to be an H.P. The Hurwitz character of $P(\lambda)$ and $Q(\lambda)$ alone does not establish the p.r. character of $Z(\lambda)$, but the p.r. character of $Z(\lambda)$ establishes the Hurwitz character of both $P(\lambda)$ and $Q(\lambda)$.

Moreover if

$$Z(\lambda) = \frac{P(\lambda)}{Q(\lambda)} = \frac{m_1 + n_1}{m_2 + n_2}$$

is a p.r. function, then not only $P(\lambda)$ and $Q(\lambda)$, but also $m_1 + n_2$ and $m_2 + n_1$ are Hurwitz polynomials.

IV. SPECIAL FORMS OF $Z(\lambda)$ IN THE TWO-ELEMENT CASES

1. LC Networks (Lossless Case)[1]

The impedance (called a *reactance function*) has the special form

$$Z(\lambda) = \frac{m_1}{n_2} \quad \text{or} \quad Z(\lambda) = \frac{n_1}{m_2} \tag{12}$$

This form may be regarded as the special case of a p.r. function whose real part is identically zero on the j-axis, that is, as the boundary case of a rational function which just barely makes the p.r. requirement. It is to be noted that in this limiting form, the p.r. function becomes the ratio of two polynomials of which one is even and the other odd.

From what is pointed out in Section III, this function may be described in either of two entirely equivalent ways. (*a*) A reactance function is one having only simple zeros and poles, alternating on the j-axis of the λ-plane. (*b*) A reactance function is one having only simple poles on the j-axis, with positive real residues.

In connection with (*a*) it is tacitly to be understood that the function is real for real λ; thus it follows that the critical frequencies occur in pairs of conjugate imaginaries, and the separation property then demands that $\lambda = 0$ and $\lambda = \infty$ be critical. The statement (*b*) is self-sufficient.

A function having these properties is realizable. For example, according to (*b*) the partial fraction expansion is guaranteed to have realizable terms, because a typical term combining a pair of conjugate poles reads

$$\frac{2k_\nu\lambda}{\lambda^2 - \lambda_\nu^2} \quad \text{with} \quad \begin{cases} k_\nu \text{ real and positive} \\ \lambda_\nu^2 \text{ real and negative} \end{cases}$$

Further detailed properties are expressed by writing

$$Z(j\omega) = jX(\omega); \quad \frac{dX}{d\omega} > 0; \quad \frac{dX}{d\omega} > \frac{X}{\omega} \text{ for } \omega > 0 \tag{13}$$

2. RC or RL Networks*[2]

In these cases the impedance $Z(\lambda)$ is not the ratio of odd and even, or of even and odd functions; the polynomials $P(\lambda)$ and $Q(\lambda)$ contain all powers of λ as they do in the general case. However, an *RL* or *RC* impedance (or admittance) is one having simple zeros and poles alternating on the negative real axis of the λ-plane.

Any rational function having this property is realizable as an *RL* or *RC* network. If the lowest critical frequency is a pole, then the function is the *impedance* of an *RC* network or the *admittance* of an *RL* network. If the lowest critical frequency is a zero, then the function is the *impedance* of an *RL* network or the *admittance* of an *RC* network.

More specifically, *for RC networks:*

$$\left(\frac{dZ}{d\lambda}\right)_{\lambda=\sigma} < 0; \quad \left(\frac{dY}{d\lambda}\right)_{\lambda=\sigma} > 0$$

At a pole of $Z(\lambda)$, or at a pole of $Y(\lambda)/\lambda$, the residue is real and positive.

For RL networks:

$$\left(\frac{dZ}{d\lambda}\right)_{\lambda=\sigma} > 0; \quad \left(\frac{dY}{d\lambda}\right)_{\lambda=\sigma} < 0$$

At a pole of $Z(\lambda)/\lambda$, or at a pole of $Y(\lambda)$, the residue is real and positive. The analogous Foster procedure in the *RC* case is to expand either $Z(\lambda)$ or $Y(\lambda)/\lambda$ into partial fractions. The analogous Foster procedure in the *RL* case is to expand either $Z(\lambda)/\lambda$ or $Y(\lambda)$ into partial fractions. From the networks thus obtained, one recognizes that for *RC* networks:

$\mathrm{Re}[Z(j\omega)]$ is continuously *decreasing* for $0 < \omega < \infty$
$\mathrm{Re}[Y(j\omega)]$ is continuously *increasing* for $0 < \omega < \infty$

while for *RL* networks the statements apply with Z and Y interchanged. Hence the minimum of $\mathrm{Re}[Z(j\omega)]$:

Occurs at $\omega = \infty$ for *RC* networks
Occurs at $\omega = 0$ for *RL* networks

and for $\mathrm{Re}[Y(j\omega)]$ the same statements apply with the references to *RC* and *RL* networks interchanged.

These results are needed for the Cauer (ladder) developments, which are accomplished through alternately removing a real part and a pole

---

[*] For detailed discussion refer to the paper by W. Cauer entitled "Die Verwirklichung von Wechselstromwiderstanden vorgeschriebener Frequenzabhängigkeit," *Arch. Elektrotech.*, **17**, 355 (1927), or Communication Networks by E. A. Guillemin, John Wiley and Sons, 1935, Chapter V.

from the given function and its inverted remainder in a continuing sequence, the subtraction being done consistently either at $\omega = 0$ or at $\omega = \infty$, with the guiding principle that each subtracted real part must be a minimum.

V. Some Remarks Relevant to the Brune Process[3]

The Brune process need not be begun until the function $Z(\lambda)$ has neither zeros nor poles for $\lambda = j\omega$, for so long as j-axis poles are contained in the function or its reciprocal, these may be removed by the Foster method just as in the reactive case. It is possible that this preliminary Foster procedure (also referred to as the "preamble" to the Brune process) ultimately leaves a remainder that is simply a constant. Otherwise one is left with a function

$$Z(\lambda) = \frac{a_0 + a_1\lambda + \cdots + a_n\lambda^n}{b_0 + b_1\lambda + \cdots + b_n\lambda^n} \tag{14}$$

in which a_0, b_0, a_n, b_n are nonzero, and neither numerator nor denominator polynomial has j-axis zeros.

The first step in the Brune process is to form $\operatorname{Re}[Z(j\omega)]$ and determine its smallest minimum (call this R_1) and the frequency $\lambda_1 = j\omega_1$ at which it occurs. The function

$$Z_1(\lambda) = Z(\lambda) - R_1 \tag{15}$$

is surely p.r. and has a pure imaginary value at $\lambda = \lambda_1$, that is

$$Z_1(\lambda_1) = jX \tag{16}$$

One then removes a series $L_1 = X/\omega_1$, leaving $Z_1(\lambda) - L_1\lambda$ which is zero at $\lambda_1 = j\omega_1$. This remainder function is p.r. only if $L_1 < 0$; that is, if $X < 0$. If, for the moment, we disregard the possibility of having $X > 0$, then it is possible next to remove a shunt branch consisting of L_2 and C_2 in series, with $L_2C_2 = -1/\lambda_1^2$. This step is indicated by

$$\frac{1}{Z_1(\lambda) - L_1\lambda} - \frac{\left(\frac{1}{L_2}\right)\lambda}{\lambda^2 - \lambda_1^2} = \frac{1}{W(\lambda)} \tag{17}$$

in which the impedance function $W(\lambda)$ is again surely p.r.

Since for $\lambda \to \infty$, $Z_1(\lambda)$ remains finite, one has

$$W(\lambda) \to \frac{-L_1L_2\lambda}{L_1 + L_2} \text{ for } \lambda \to \infty \tag{18}$$

Removing from $W(\lambda)$ a series inductance

$$L_3 = \frac{-L_1L_2}{L_1 + L_2} \tag{19}$$

leaves the function

$$Z_2(\lambda) = W(\lambda) - L_3\lambda \tag{20}$$

having no j-axis poles because $1/W$ has no finite j-axis zeros since the real part of $Z_1(j\omega)$ would there have to be zero. $Z_2(\lambda)$ has the same form as $Z(\lambda)$ given by eq. 14, but is simpler in that the polynomials have the degree $n - 2$.

This cycle of steps leads to the network of Fig. 1 in which the parameters of the reactive two terminal-pair network have values given by

$$L_1 = \frac{Z_1(\lambda_1)}{\lambda_1} < 0; \quad L_2 = \frac{1}{2}\left[\left(\frac{dZ_1}{d\lambda}\right)_{\lambda=\lambda_1} - L_1\right] > 0$$
$$L_2C_2 = -1/\lambda_1^2; \quad L_3 = -L_1L_2/(L_1 + L_2) > 0 \tag{21}$$

Now if $X > 0$, an entirely analogous procedure on an admittance basis is possible. Namely, one removes from $1/Z_1(\lambda)$ a shunt capacitance

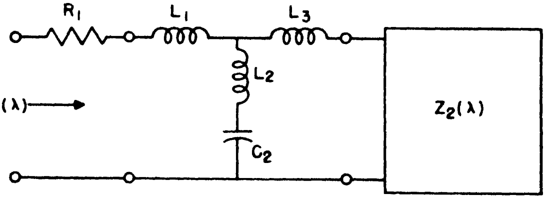


FIG. 1 Network resulting from one Brune cycle carried out on the impedance basis (appropriate to $X < 0$)

$C_1' = 1/\lambda_1 Z_1(\lambda_1)$ leaving a p.r. remainder

$$\frac{1}{Z_1(\lambda)} - C_1'\lambda \tag{22}$$

which is zero at $\lambda = \lambda_1$. Hence one can next remove a series branch consisting of L_2' and C_2' in parallel, with $L_2'C_2' = -1/\lambda_1^2$. This step is indicated by

$$\frac{1}{\dfrac{1}{Z_1(\lambda)} - C_1'\lambda} - \frac{\left(\dfrac{1}{C_2'}\right)\lambda}{\lambda^2 - \lambda_1^2} = \frac{1}{W(\lambda)} \tag{23}$$

in which the admittance function $W(\lambda)$ is again surely p.r.

Since for $\lambda \to \infty$, $1/Z_1(\lambda)$ remains finite, one has

$$W(\lambda) \to \frac{-C_1'C_2'\lambda}{C_1' + C_2'} \text{ for } \lambda \to \infty \tag{24}$$

Removing from $W(\lambda)$ a shunt capacitance

$$C_3' = \frac{-C_1'C_2'}{C_1' + C_2'} \tag{25}$$

leaves the function

$$\frac{1}{Z_2(\lambda)} = W(\lambda) - C_3'\lambda \tag{26}$$

having no j-axis poles because $1/W$ has no finite j-axis zeros since the real part of $1/Z_1(j\omega)$ would there have to be zero. Again $Z_2(\lambda)$ is simpler than $Z_1(\lambda)$ but has the same form.

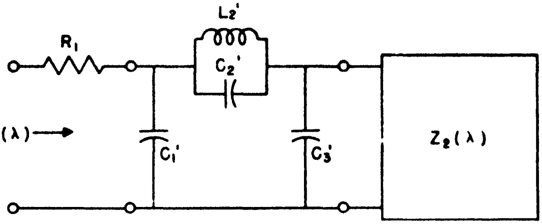


FIG 2 Network resulting from one Brune cycle carried out on the admittance basis (appropriate to $X > 0$)

This alternate cycle of steps leads to the network of Fig. 2 in which the parameters of the reactive two terminal-pair network have values given by

$$C_1' = 1/\lambda_1 Z_1(\lambda_1) < 0; \quad C_2' = \frac{1}{2}\left[\frac{-1}{Z_1^2(\lambda_1)}\left(\frac{dZ_1}{d\lambda}\right)_{\lambda=\lambda_1} - C_1'\right] > 0 \tag{27}$$
$$L_2'C_2' = -1/\lambda_1^2; \quad C_3' = -C_1'C_2'/(C_1' + C_2') > 0$$

While the negativeness of L_1 in the network of Fig. 1 can be taken care of through transforming the T of inductances into a pair of mutually coupled coils (with unity coupling), an analogous process of overcoming the practical objection to the negativeness of C_1' in the network of Fig. 2 does not exist. However, if one substitutes from eqs. 21 into 27 it is found that*

---

* The reader should note well that the symbols L_1, L_2, L_3, C_2 appearing in the following relations for the primed parameters are at this point in the argument merely abbreviations for the more cumbersome expressions for these symbols in terms of $Z_1(\lambda)$, λ_1, $(dZ_1/d\lambda)$, etc., as given by eqs. 21. That they may be identified with the correspondingly denoted parameter values in Fig. 1 is the result which the present manipulations are leading up to. Unless the reader sees this distinction clearly, the following derivations appear to be pointless and the ultimate conclusions unconvincing.

$$C_1' = \frac{1}{L_1\lambda_1^2} = \frac{-L_2C_2}{L_1} \tag{28}$$

$$C_2' = \frac{1}{2}\left[\frac{-1}{L_1^2\lambda_1^2}(2L_2 + L_1) - \frac{L_1}{L_1^2\lambda_1^2}\right] = \frac{-(L_1 + L_2)}{L_1^2\lambda_1^2} = \frac{L_2C_2}{L_1}\left(1 + \frac{L_2}{L_1}\right) \tag{29}$$

$$C_3' = C_2\left(1 + \frac{L_2}{L_1}\right); \quad C_1' + C_2' = \left(\frac{L_2}{L_1}\right)^2 C_2$$

$$C_2' + C_3' = \left(1 + \frac{L_2}{L_1}\right)^2 C_2; \quad C_1' + C_3' = C_2 \tag{30}$$

$$\frac{1}{L_2'} = \frac{C_2'}{L_2C_2} = \frac{1}{L_1}\left(1 + \frac{L_2}{L_1}\right)$$

Hence the y-system for the network of Fig. 2 is

$$y_{11} = \left(\frac{L_2}{L_1}\right)^2 C_2\lambda + \left(1 + \frac{L_2}{L_1}\right)\frac{1}{L_1\lambda}; \quad y_{22} = \left(1 + \frac{L_2}{L_1}\right)^2 C_2\lambda + \left(1 + \frac{L_2}{L_1}\right)\frac{1}{L_1\lambda} \tag{31}$$

$$y_{12} = -\left(\frac{L_2}{L_1}\right)\left(1 + \frac{L_2}{L_1}\right)C_2\lambda - \left(1 + \frac{L_2}{L_1}\right)\frac{1}{L_1\lambda}; \quad |y| = \frac{C_1' + C_3'}{L_2'} = \frac{C_2}{L_1}\left(1 + \frac{L_2}{L_1}\right)$$

Here $|y|$ denotes the determinant $y_{11}y_{22} - y_{12}^2$ of the y-system.

Converting to the equivalent z-system yields

$$z_{11} = \frac{y_{22}}{|y|} = \left(1 + \frac{L_2}{L_1}\right)L_1\lambda + \frac{1}{C_2\lambda} = (L_1 + L_2)\lambda + \frac{1}{C_2\lambda} \tag{32}$$

$$z_{22} = \frac{y_{11}}{|y|} = \frac{(L_2/L_1)^2L_1\lambda}{\left(1 + \dfrac{L_2}{L_1}\right)} + \frac{1}{C_2\lambda} = (L_2 + L_3)\lambda + \frac{1}{C_2\lambda} \tag{33}$$

$$z_{12} = L_2\lambda + \frac{1}{C_2\lambda} \tag{34}$$

In connection with the second step in eq. 33 note that the close coupling condition among the L's gives

$$(L_2 + L_3)(L_1 + L_2) = L_2^2$$

or

$$\frac{L_2 + L_3}{L_1}\left(1 + \frac{L_2}{L_1}\right) = \left(\frac{L_2}{L_1}\right)^2 \tag{35}$$

We observe that the z's given by eqs. 32, 33, 34 are those for the reactive two terminal-pair network of Fig. 1. Therefore, the two networks of Figs. 1 and 2 are demonstrated to be equivalent; and it is also demonstrated that the same values for the parameters of the network in Fig. 1 are obtained whether they are computed from the relations 21 directly or obtained through the conversion of the network of Fig. 2 into that of Fig. 1, having first found the parameter values in the former network from the relations 27. Stated in a different way: It is not necessary to use the alternate procedure on an admittance basis when X in eq. 16 is positive in spite of the fact that the function $Z_1(\lambda) - L_1\lambda$ then is no longer p.r., for if one nevertheless determines the network of Fig. 1 from the relations 21, all results are identical with those obtained on the alternate admittance basis followed by a subsequent conversion to the network of the impedance basis

VI THE DARLINGTON PROCEDURE[1] FOR THE SOLUTION OF THE BRUNE PROBLEM SKELETONIZED

One of the most significant contributions contained in Darlington's work shows that any p.r. driving point function is realizable as the input impedance to a lossless two terminal-pair network terminated in a pure resistance, Fig. 3 The procedure for finding the network N constitutes not only an alternate synthesis method to that of Brune but also, what is far more significant, it forms the basis for synthesis of lossless two terminal-pair networks with resistive terminations for prescribed transfer characteristics.

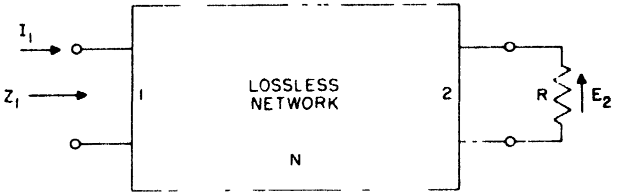


FIG. 3 Physical realization for a positive real driving point impedance function according to Darlington The associated transfer impedance is $Z_{12} = E_2/I_1$.

The first step in the discussion of Darlington's alternate method of solving the Brune problem is to express the input impedance Z_1 of Fig. 3 as

$$Z_1 = \frac{(z_{11}z_{22} - z_{12}^2) + z_{11}R}{z_{22} + R} \tag{36}$$

in which z_{11}, z_{22}, z_{12} is the set of open-circuit driving point and transfer impedances of the two terminal-pair network, as usually defined. The function $(z_{11}z_{22} - z_{12}^2)$ is the determinant of the matrix formed with these z_{ik}'s, and is conveniently abbreviated as $|z|$. The elements of the inverse matrix (which are the familiar short-circuit driving point and transfer admittances of the network N) are then expressible as

$$y_{11} = \frac{z_{22}}{|z|}; \quad y_{22} = \frac{z_{11}}{|z|}; \quad y_{12} = -\frac{z_{12}}{|z|} \tag{37}$$

Thus eq. 36 may alternatively be written

$$Z_1 = z_{11} \times \frac{\dfrac{1}{y_{22}} + R}{z_{22} + R} \tag{38}$$

while, according to eq. 1, one has

$$Z_1 = \frac{P(\lambda)}{Q(\lambda)} = \frac{m_1 + n_1}{m_2 + n_2} \tag{39}$$

Remembering that the ratio of any even part to any odd part of the polynomials $P(\lambda)$ and $Q(\lambda)$ yields a physically realizable reactance or susceptance function, the following manipulations quite naturally suggest themselves

$$Z_1 = \frac{m_1}{n_2} \times \frac{(n_1/m_1) + 1}{(m_2/n_2) + 1} \tag{40}$$

or

$$Z_1 = \frac{n_1}{m_2} \times \frac{(m_1/n_1) + 1}{(n_2/m_2) + 1} \tag{41}$$

for, with the normalization $R = 1$ ohm, one has through comparing eq. 38 with 40 (case A) the set of identifications

$$z_{11} = \frac{m_1}{n_2}; \quad z_{22} = \frac{m_2}{n_2}; \quad y_{22} = \frac{m_1}{n_1} \tag{42}$$

or through comparing eq. 38 with 41 (case B) the alternate set of identifications

$$z_{11} = \frac{n_1}{m_2}; \quad z_{22} = \frac{n_2}{m_2}; \quad y_{22} = \frac{n_1}{m_1} \tag{43}$$

Although the realizability of each of these driving point functions separately is assured, it is not apparent that collectively they represent a realizable two terminal-pair network. In this regard observe first that eqs. 37 together with 42 or 43 yields

$$z_{11}z_{22} - z_{12}^2 = \frac{z_{11}}{y_{22}} = \frac{n_1}{n_2} \quad \text{(case A)} \qquad \qquad (44)$$
$$= \frac{m_1}{m_2} \quad \text{(case B)}$$

Thus we have

$$z_{12} = \frac{\sqrt{m_1m_2 - n_1n_2}}{n_2} \quad \text{(case A)} \qquad (45)$$

and

$$z_{12} = \frac{\sqrt{n_1n_2 - m_1m_2}}{m_2} \quad \text{(case B)} \qquad (46)$$

With the use of 4 one has

$$z_{12} = \frac{\sqrt{A(-\lambda^2)}}{n_2} \quad \text{(case A)} \qquad \qquad (47)$$
$$z_{12} = \frac{\sqrt{-A(-\lambda^2)}}{m_2} \quad \text{(case B)}$$

Since z_{12} must be a rational function it is necessary that the polynomial $\pm A(-\lambda^2)$ be a full square, which is the same as saying that its λ^2-zeros (placed in evidence by the form given in eqs. 5) all be of even multiplicity. According to the discussion in Section II, the p.r. character of Z_1 assures that all negative real λ^2-zeros be of even multiplicity, but the other types of zeros need not be. This eventuality is met through multiplying numerator and denominator in the expression 39 for Z_1 by an appropriate auxiliary Hurwitz polynomial $P_0 = m_0 + n_0$, thus

$$Z_1 = \frac{P(\lambda)P_0(\lambda)}{Q(\lambda)P_0(\lambda)} = \frac{(m_1m_0 + n_1n_0) + (m_1n_0 + n_1m_0)}{(m_2m_0 + n_2n_0) + (m_2n_0 + n_2m_0)} \qquad (48)$$

which is a trivial operation so far as Z_1 is concerned but, as a straightforward algebraic calculation shows, it leads to the revised function

$$\pm A(-\lambda^2) = \pm(m_1m_2 - n_1n_2)(m_0^2 - n_0^2) \qquad (49)$$

Through choosing $(m_0^2 - n_0^2)$ equal to those factors in the original $A(-\lambda^2)$ function as given by eqs. 5 which occur with odd multiplicity, one obtains a revised function $A(-\lambda^2)$ that is a full square. It also becomes clear that if the original $A(-\lambda^2)$ function has a simple zero root (for example, if $\lambda_1^2 = 0$ in eqs. 5) then, assuming that the remaining factors are at least quadratic, it is $-A(-\lambda^2)$ that is positive and a full square. In this case $A(-\lambda^2)$ is an odd function of λ^2, and it becomes clear that cases A and B, as distinguished above, result according to whether the revised $A(-\lambda^2)$ polynomial is even or odd respectively* as a function of λ^2. It may additionally be pointed out that $A(-\lambda^2)$ can always be made even through an appropriate choice of the factors comprising $(m_0^2 - n_0^2)$, but only at the expense of ultimately yielding z_{12} of higher degree and consequently obtaining more elements in the resulting network.

The polynomial $P_0 = (m_0 + n_0)$ is found from the chosen $(m_0^2 - n_0^2)$ through observing that

$$m_0^2 - n_0^2 = (m_0 + n_0)(m_0 - n_0) \tag{50}$$

and noting that the zeros of $(m_0 - n_0)$ are those of $m_0 + n_0$ reflected about the j-axis. Since P_0 must be Hurwitz, the process of forming $m_0 + n_0$ from a given $m_0^2 - n_0^2$ is clearly that of constructing a polynomial out of the left half-plane zeros of $m_0^2 - n_0^2$.

Throughout the further discussion of Darlington's solution to the Brune problem it will be assumed that the process just described for rendering $\sqrt{\pm(m_1m_2 - n_1n_2)}$ an ordinary polynomial (real for real λ) has been carried out previous to our consideration of the given impedance Z_1, so that when we arrive at the expressions 45, 46, or 47 for z_{12} the question of a possible irrational character does not arise. It then merely remains to show that the set of three functions z_{11}, z_{22}, z_{12} represent a realizable lossless two terminal-pair network. This will be the case if the so-called *residue condition*

$$k_{11}k_{22} - k_{12}^2 \geq 0 \tag{51}$$

is fulfilled. Here k_{11}, k_{22}, and k_{12} are respectively the residues of z_{11}, z_{22}, and z_{12} at any j-axis poles which these functions possess. It must be shown, therefore, that this residue condition is fulfilled for a set of z's resulting from relations 42 and 45, or 43 and 46.

Considering case A (eqs. 42 and 45), and noting that a pole corresponds to a zero of n_2, we see (according to a common procedure for the evaluation of residues of rational functions) that

$$k_{11} = \left(\frac{m_1}{n_2'}\right)_{\lambda=\lambda_\nu}; \quad k_{22} = \left(\frac{m_2}{n_2'}\right)_{\lambda=\lambda_\nu}; \quad k_{12} = \left(\frac{\sqrt{m_1m_2}}{n_2'}\right)_{\lambda=\lambda_\nu} \tag{52}$$

in which the prime indicates differentiation and λ_ν is the j-axis pole in question. It follows that the residue condition 51 is fulfilled with the equals sign. Thus, for any given p.r. function Z_1, a set of impedances z_{11}, z_{22}, z_{12} leading to a realizable lossless two terminal-pair network can always be found.

---

* In either case the function z_{12} is seen to be the ratio of two polynomials of which one is even and the other odd, as required by the lossless character of the network N.

VII. SYNTHESIS OF THE SINGLE-LOADED LOSSLESS COUPLING NETWORK FOR A PRESCRIBED MAGNITUDE OF TRANSFER IMPEDANCE

First of all it should be recalled that the transfer function (impedance, admittance, or dimensionless ratio) of a physically realizable network is not required to be a p.r. function. Any rational function, regular at $\lambda = \infty$, having its poles restricted to the left half-plane, is acceptable. If resistances are entirely absent (in the load impedances as well as in the coupling network itself) then the transfer function is permitted to have simple poles on the j-axis (including the points $\lambda = 0$ and $\lambda = \infty$), but such a situation is not frequently met in practical problems.

With reference to Fig. 3, the transfer impedance of the lossless network N loaded by the single resistance R is defined as

$$Z_{12} = E_2/I_1 \tag{53}$$

In the sinusoidal steady state, the average power input must equal that delivered to the load; hence

$$|I_1|^2 \times \mathrm{Re}[Z_1(j\omega)] = |E_2|^2/R \tag{54}$$

or

$$R \times \mathrm{Re}[Z_1(j\omega)] = \left|\frac{E_2}{I_1}\right|^2 = |Z_{12}(j\omega)|^2 \tag{55}$$

If the load resistance R is normalized at 1 ohm, one may say that the squared magnitude of the transfer impedance $Z_{12}(j\omega)$ numerically equals the real part of the driving point impedance $Z_1(j\omega)$. According to eq. 3 one may write

$$|Z_{12}(j\omega)|^2 = \frac{A_0 + A_1\omega^2 + \cdots + A_n\omega^{2n}}{B_0 + B_1\omega^2 + \cdots + B_n\omega^{2n}} = \frac{A(\omega^2)}{B(\omega^2)} \tag{56}$$

As was first shown by Gewertz,[5] one may through an algebraic process construct from this real part of $Z_1(j\omega)$ the driving point impedance $Z_1(\lambda)$. First it should be observed from eqs. 4 that

$$B(-\lambda^2) = m_2^2 - n_2^2 = (m_2 + n_2)(m_2 - n_2) \tag{57}$$

may be used to construct the polynomial $Q(\lambda) = m_2 + n_2$ just as $P_0 = m_0 + n_0$ is shown in the previous article to be constructible from $m_0^2 - n_0^2$. Except for an unimportant constant multiplier, $Q(\lambda)$ is thus formed from the left half-plane zeros of $m_2^2 - n_2^2$.

Next, following the notation in eqs. 1 and 3, it is observed that

$$\begin{aligned} A_0 + A_1\omega^2 + \cdots + A_n\omega^{2n} &= (m_1m_2 - n_1n_2)_{\lambda=j\omega} \\ &= (a_0 - a_2\omega^2 + a_4\omega^4 - \cdots)(b_0 - b_2\omega^2 + b_4\omega^4 - \cdots) \\ &\quad + \omega^2(a_1 - a_3\omega^2 + a_5\omega^4 - \cdots)(b_1 - b_3\omega^2 + b_5\omega^4 - \cdots) \\ &= a_0b_0 + (-a_0b_2 + a_1b_1 - a_2b_0)\omega^2 \\ &\qquad + (a_0b_4 - a_1b_3 + a_2b_2 - a_3b_1 + a_4b_0)\omega^4 + \cdots \end{aligned} \tag{58}$$

Equating coefficients of like powers of ω^2 gives

$$\begin{aligned} A_0 &= a_0 b_0 \\ A_1 &= -a_0 b_2 + a_1 b_1 - a_2 b_0 \\ A_2 &= a_0 b_4 - a_1 b_3 + a_2 b_2 - a_3 b_1 + a_4 b_0 \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \end{aligned} \tag{59}$$

yielding the general formula

$$A_r = \sum_{s=-r}^{s=r} a_{r+s} b_{r-s} \times (-1)^s; \quad r = 0, 1, \cdots n \tag{60}$$

Coefficients a_k or b_k are, of course, zero for $k > n$.

Equations 59 may be solved for the coefficients $a_0 \cdots a_n$ of the polynomial $P(\lambda)$ the numerator of $Z_1(\lambda)$ in terms of the known coefficients $A_0 \cdots A_n$ and the coefficients $b_0 \cdots b_n$ of $Q(\lambda)$ previously determined.

An alternative method of determining $Z_1(\lambda)$ from $\mathrm{Re}[Z_1(j\omega)]$, due to Bode,[6] is the following. One recognizes readily that

$$\mathrm{Re}[Z_1(j\omega)] = \tfrac{1}{2}[Z_1(\lambda) + Z_1(-\lambda)]_{\lambda=j\omega} \tag{61}$$

Hence

$$\frac{2A(-\lambda^2)}{B(-\lambda^2)} = Z_1(\lambda) + Z_1(-\lambda) \tag{62}$$

The constellation of poles of $Z_1(\lambda)$ lies in the left half-plane; its image about the j-axis is that of $Z_1(-\lambda)$. The function $2A(-\lambda^2)/B(-\lambda^2) = f(\lambda)$ has both of these pole constellations. The residue of $f(\lambda)$ in one of its left half-plane poles is the same as the residue of $Z_1(\lambda)$ in the corresponding pole (since the residue of $Z_1(-\lambda)$ is there zero). Since for $\lambda \to \infty$, $Z_1(\lambda) \to Z_1(-\lambda) \to R$ (a constant), the partial fraction expansion of $Z_1(\lambda)$ is given by R plus the principal parts of Laurent expansions of the rational function $f(\lambda)$ in its left half-plane poles.

If the partial fraction expansion of $Z_1(\lambda)$, rather than its representation as a quotient of polynomials, is wanted, Bode's procedure is more direct than that of Gewertz. If the polynomial form for $Z_1(\lambda)$ is preferred, the Gewertz method is computationally shorter.

VIII. CAUER'S METHOD OF SYNTHESIS FROM[7] A SPECIFIED $|Z_{12}(j\omega)|^2$

A more direct method of solving the problem discussed in the previous section was given by Cauer. His method begins by recognizing that, in a straightforward manner (preferably using Thévenin's theorem), one may express the transfer impedance $Z_{12}(\lambda)$ for a 1-ohm load as

$$Z_{12} = \frac{z_{12}}{1 + z_{22}} = \frac{g(\lambda)}{h(\lambda)} \tag{63}$$

Here $h(\lambda)$ is a Hurwitz polynomial (stability requirement) and so contains both even and odd powers of λ, while the polynomial $g(\lambda)$ is either even or odd since z_{12} (like a driving point reactance function) must be the ratio of two polynomials of which one is even and the other odd.

If one writes

$$h(\lambda) = m + n \tag{64}$$

in which m is even and n odd, it is seen that

$$|Z_{12}(j\omega)|^2 = \left[\frac{\pm g^2(\lambda)}{m^2 - n^2}\right]_{\lambda=j\omega} = \frac{A(\omega^2)}{B(\omega^2)} \tag{65}$$

the $\pm$ signs corresponding respectively to $g(\lambda)$ being even or odd. Thus

$$\begin{aligned} A(-\lambda^2) &= \pm g^2(\lambda) \\ B(-\lambda^2) &= m^2 - n^2 \end{aligned} \tag{66}$$

Again $A(-\lambda^2)$ must be a full square. If it is not a full square at the outset, one multiplies numerator and denominator in the given expression for $|Z_{12}(j\omega)|^2$ by appropriate identical factors so as to bring about this condition. Upon subsequently writing the relations 66, one obtains the polynomial $g(\lambda)$ at once as the square root of $A(-\lambda^2)$, and the polynomial $h(\lambda) = m + n$ is then constructed through use of the left half-plane zeros of $B(-\lambda^2)$ in the manner discussed previously.

Returning now to the eq. 63, one may write

$$Z_{12} = \frac{g}{m + n} = \frac{g/n}{1 + (m/n)} = \frac{g/m}{1 + (n/m)} \tag{67}$$

and make either of the identifications

$$z_{12} = (g/n); \quad z_{22} = (m/n) \tag{68}$$

or

$$z_{12} = (g/m); \quad z_{22} = (n/m) \tag{69}$$

according to whether g is even or odd respectively.

Thus one has found a pair of acceptable functions z_{12} and z_{22} so far as the realizability of a corresponding lossless two terminal-pair network is concerned. In order to be able to carry out the synthesis of this network in the usual manner, which proceeds from the partial fraction expansions of the three functions z_{11}, z_{22}, z_{12}, the completion of this set through the association of an appropriate z_{11} with the functions z_{22} and z_{12} found from eq. 68 or 69 must first be accomplished.

This step is readily carried out through use of the residue condition expressed by eq. 51. Thus, after the partial fraction expansions of z_{22} and z_{12} are written down (the coefficients in the terms of these expansions are the residues k_{22} and k_{12}), it is a simple and straightforward matter to write the partial fraction expansion for an appropriate z_{11}-function since the residues k_{11} are any values satisfying the condition 51. One has the choice of fulfilling this condition with the inequality or with the equality sign, and so the question arises as to the significance or implication of either procedure.

Clarification on this point is had through first recalling the expression 36 for the input impedance. The z-determinant $(z_{11}z_{22} - z_{12}^2)$ appearing here should be visualized as computed through substituting the partial fraction expansions for the z's. Careful reflection (or better still, the writing out of a simple example) reveals that the result does not contain terms representing second order poles if the residue condition 51 is fulfilled with the equals sign at each pole, while second order poles in the function $(z_{11}z_{22} - z_{12}^2)$ are surely present if the inequality in 51 holds at one or more poles. One may say that if the residue condition is fulfilled with the equals sign at all poles, then the z-determinant has only simple poles; and the converse of this statement is also true. Under these circumstances one observes from eq 36 that the input impedance Z_1 does not contain the j-axis poles of z_{11}, z_{22}, z_{12} at all since numerator and denominator are both of first order However, at any pole where the residue condition 51 is fulfilled with the inequality, there Z_1 will surely contain that pole also, because $(z_{11}, z_{22} - z_{12}^2)$ will there have a second order pole so that the numerator of eq 36 will be one order higher than the denominator.

The conclusion is that the results for z_{22} and z_{12} expressed by eqs. 68 and 69 suffice to determine the lossless two terminal-pair network since one can readily associate an appropriate z_{11}-function through use of the residue condition 51. Moreover, if this condition is written with the equals sign, then the result yields an input impedance Z_1 having no j-axis poles (this is referred to as a *minimum reactive* driving point impedance). The desired transfer impedance stays the same whether the residue condition is met so as to yield a minimum reactive Z_1 or not.

Speaking of the minimum reactive character of Z_1 reminds one of the question regarding the minimum phase or nonminimum phase character of the transfer impedance Z_{12}. Here it will be recalled that Z_{12} is minimum phase if (and only if) its zeros as well as its poles lie in the left-half λ-plane. These zeros are those of the polynomial $g(\lambda)$ in eq. 63.

Since $g(\lambda)$ is either even or odd, its zeros (except for one at $\lambda = 0$ if $g(\lambda)$ is odd) occur either as pairs of real values or as quadruplets, spaced symmetrically about the real and imaginary axes of the λ-plane. In general, therefore, there are always some zeros in the right half-plane, and so the resulting Z_{12} is nonminimum phase.

This result is a consequence of stipulating that the network N shall be lossless, for then z_{12} in eq. 63 must be the ratio of two polynomials of which one is even and the other odd, thus yielding a $g(\lambda)$ that is either even or odd. There is one notable exception to this conclusion. Namely, if *all* the zeros of Z_{12} (or z_{12}) fall upon the j-axis, they are appropriately interpreted as belonging to the left half-plane (because the inevitable incidental dissipation present in an actual physical realization will place them there in spite of their being on the j-axis theoretically), and in a limiting sense Z_{12} then becomes minimum phase. This case is important practically because most filter designs are carried out (for reasons of economy) by choosing all the zeros of Z_{12} on the j-axis

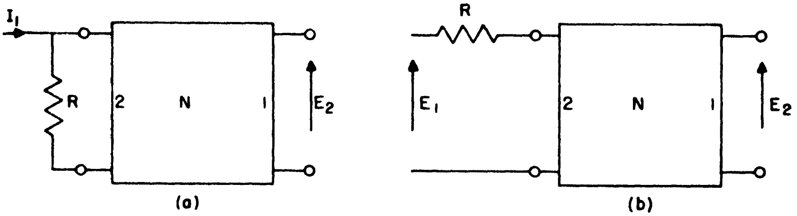
IX. COMPLEMENTARY IMPEDANCES; CONSTANT-RESISTANCE FILTER GROUPS

It is well at this point to pause for a moment and reflect upon what has been accomplished with regard to the general problem of designing coupling networks for prescribed transfer characteristics, and what remains undone. In this regard the results so far may be adapted to meet various needs. As the network of Fig. 3 stands it may, for example, be used where the input is the plate current of a pentode. Except for a constant multiplier, the ratio of E_2 to the voltage at the grid of the pentode is given by Z_{12}, since the plate current is essentially proportional to the grid voltage. This is a situation where the source impedance is very large compared with Z_1.

If the reciprocity theorem is applied to the situation given in Fig. 3, we arrive at the one shown in Fig. 4*a*, which, through a current-to-voltage source conversion, yields the arrangement in Fig. 4*b*. The voltage ratio E_2/E_1 is again proportional to Z_{12} as considered above. Thus the design is adaptable to a situation where the source has finite resistance but the load is essentially an open circuit (as, for example, the grid terminals to an amplifier tube).

Further adaptation of the present result to other practical situations is had through use of the duality principle. Thus if we change the situation of Fig. 3 to one in which the input is a voltage E_1 applied to the terminal pair 1 of the lossless network N and E_2 is replaced by the current I_2 through the load R, and we use the letter Y or y in place of Z or z (impedance functions become admittance functions), everything that has been said remains intact. We merely shift from an impedance basis to the dual admittance basis. Instead of designing for a transfer impedance $Z_{12} = E_2/I_1$, we design for a transfer admittance $Y_{12} = I_2/E_1$. Since the voltage across the load resistance R is proportional to I_2, the ratio of load voltage to input voltage is simply a constant times Y_{12}. Our method of design is now appropriate where the source impedance is negligible compared with the input impedance to the lossless network.

In all these applications one of the associated external impedances (source or load) is either very large or very small compared with Z_1 (Fig. 3). It is necessary to be able to meet situations where such restrictions do not apply; that is, where finite external resistances are associated with the lossless network at both input and output ends.



(a) (b)

FIG. 4 Alternate interpretation of the transfer impedance appropriate to Fig 3 through use of the reciprocity theorem followed by a source conversion.

One method of meeting this need is to place in series with the input terminals (Fig. 3) an impedance Z_1^c such that $Z_1 + Z_1^c$ equals a constant, for then the presence of a source with finite internal resistance leaves I_1 proportional to the source voltage, and all is well. The network realizing Z_1^c is called the complement of N; Z_1 and Z_1^c are referred to as complementary impedances.*

The essential condition for the existence of such a complementary network is that Z_1 be minimum reactive since it would clearly be impossible to have $Z_1 + Z_1^c$ equal a constant if Z_1 had j-axis poles, for there is no way in which Z_1^c could cancel these and be realizable by a passive network. With Z_1 minimum reactive, the imaginary parts of Z_1 and Z_1^c are capable of canceling each other while the real parts are so related that their sum is a constant (the real parts are complementary). Because of the implicit relationship between the real and imaginary parts of a minimum reactive impedance, it is sufficient to find a Z_1^c whose real part is complementary to that of Z_1; the imaginary parts will then automatically cancel.

---

* Here the network N inclusive of the terminal resistance is meant.

Either of two procedures may be followed to find a network for Z_1^c. In one of these the complementary impedance is formed directly through writing $Z_1^c = K - Z_1$ in which K is a positive real constant at least as large as the largest value of the real part of $Z_1(j\omega)$. The function Z_1^c thus found is surely p.r. if Z_1 is p.r. and minimum reactive. This p.r. Z_1^c-function may then be synthesized either by Brune's or Darlington's method.

Alternatively one begins with

$$\mathrm{Re}[Z_1^c(j\omega)] = K - \mathrm{Re}[Z_1(j\omega)] \tag{70}$$

and, using eq. 55 with $R = 1$, obtain

$$\mathrm{Re}[Z_1^c(j\omega)] = K - |Z_{12}(j\omega)|^2 = |Z_{12}^c(j\omega)|^2 \tag{71}$$

in which Z_{12}^c is a transfer impedance for the desired complementary network which, like the given network N, is likewise regarded as realized by a lossless two terminal-pair network terminated in a 1-ohm resistance. From the function Z_{12}^c one synthesizes a network according to methods already described so as to yield a minimum reactive driving point impedance; the latter is the desired Z_1^c

The complementary network thus obtained may be regarded simply as an impedance-correcting network in the sense that, in combination with Z_1, it "corrects" the latter so as to make it constant. Since this network has a transfer function (according to eq. 71) that is complementary to that of the given network, we may regard each of the two networks as having an independent status, so to speak, but with the networks so paired that they mutually complement each other. Thus if the given network is a low-pass filter, for example, the second is a high-pass filter; and the combination has an input impedance equal to a constant. In this light the structure as a whole is spoken of as a *constant-resistance filter group.*

Through an obvious shift to the dual admittance basis one obtains the filter group in the form of networks whose inputs appear in parallel instead of being in series.

X. ANOTHER WAY OF DESIGNING FOR FINITE RESISTANCES AT BOTH SOURCE AND LOAD[8]

Consider the situation shown in Fig. 5 where the lossless coupling network has associated external resistances at both ends. Write $Z_1(j\omega) = R + jX$.

The power entering the network N must equal that leaving it, so

$$|I_1|^2R = |E_2|^2/R_2 \tag{72}$$

Since

$$I_1/I_{10} = R_1/(R_1 + Z_1) \tag{73}$$

we have

$$|I_1|^2 = |I_{10}|^2R_1^2/|R_1 + Z_1|^2 \tag{74}$$

Together with 72 this gives

$$\frac{|E_2|^2}{R_1R_2|I_{10}|^2} = \frac{RR_1}{|R_1 + Z_1|^2} = \frac{1}{4}\left(1 - \left|\frac{R_1 - Z_1}{R_1 + \bar{Z}_1}\right|^2\right) \tag{75}$$

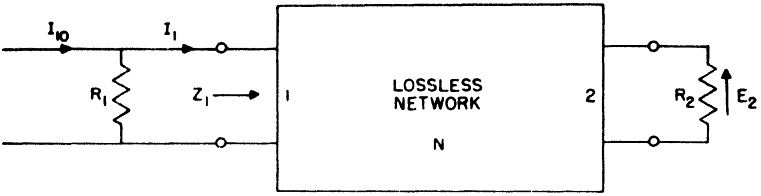


FIG 5 Where the lossless coupling network has resistive loading at both ends, the associated transfer impedance is the function E_2/I_{10}

If the current source I_{10} paralleled by R_1 is converted into an equivalent voltage source of the value $E_1 = R_1I_{10}$, we find

$$\frac{4|E_2|^2}{R_1R_2|I_{10}|^2} = \frac{|E_2|^2}{|E_1|^2}\frac{R_2}{4R_1} = |t|^2 \tag{76}$$

This quantity is evidently the ratio of power delivered to the load R_2 to the maximum power deliverable by the source. Evidently $|t|^2$ (the squared absolute value of a *transmission coefficient* t) can at most equal unity.

The function

$$\rho = \frac{R_1 - Z_1}{R_1 + Z_1} \tag{77}$$

is recognized as the *reflection coefficient* at the input terminal pair of the lossless network. In terms of $|t|^2$ and $|\rho|^2$, eq. 75 states the logical fact that whatever power is not deliverable to the load must be returned to the source, that is

$$|t|^2 = 1 - |\rho|^2 \tag{78}$$

The given function in a situation of this sort presumably would be

$$|t|^2 = \frac{C(\omega^2)}{B(\omega^2)} \tag{79}$$

a quotient of polynomials in ω^2. From this function one readily obtains

$$|\rho|^2 = \frac{B(\omega^2) - C(\omega^2)}{B(\omega^2)} = \frac{D(\omega^2)}{B(\omega^2)} \tag{80}$$

If we write

$$\rho(\lambda) = \frac{p(\lambda)}{q(\lambda)} \tag{81}$$

then

$$\frac{D(-\lambda^2)}{B(-\lambda^2)} = \frac{p(\lambda) \cdot p(-\lambda)}{q(\lambda) \cdot q(-\lambda)} \tag{82}$$

Since the poles of $\rho(\lambda)$ must lie in the left half-plane (this is evident from eq. 77), it is clear that $q(\lambda)$ is formed from the left half-plane zeros of $B(-\lambda^2)$, following a well-established pattern. In precisely the same manner $p(\lambda)$ may be constructed from the left half-plane zeros of $D(-\lambda^2)$, although one is here permitted to vary the procedure through picking some or even all of the zeros out of the right half-plane, since the zeros of $\rho(\lambda)$ are not restricted. (It may be pointed out here, however, that restriction of the zeros of ρ to the left half-plane is a necessary condition in the process of maximizing the gain-bandwidth product for a given associated shunt capacitance; see H. W. Bode, Network Analysis and Feedback Amplifier Design, D. Van Nostrand, 1945, pp. 360–368.)

With the function $\rho(\lambda)$ determined, one has from eq. 77

$$\frac{Z_1}{R_1} = \frac{P(\lambda)}{Q(\lambda)} = \frac{m_1 + n_1}{m_2 + n_2} = \frac{1 - \rho}{1 + \rho} \tag{83}$$

and the discussion of previous articles may then be applied to find the network N. It is also helpful in this connection to note that

$$\rho(\lambda) = \frac{(m_2 - m_1) + (n_2 - n_1)}{(m_2 + m_1) + (n_2 + n_1)} \tag{84}$$

and with the use of eq. 80,

$$\frac{D(-\lambda^2)}{B(-\lambda^2)} = \frac{(m_2 - m_1)^2 - (n_2 - n_1)^2}{(m_2 + m_1)^2 - (n_2 + n_1)^2} \tag{85}$$

Thus one can obtain m_1, m_2, n_1, n_2 directly, and hence the z_{11}, z_{22}, z_{12}, without bothering to form Z_1.

XI. THE CONSTANT-RESISTANCE LATTICE

The foregoing discussion of the problem of synthesis for a prescribed transfer function is inadequate principally because it does not afford a method of obtaining the prescribed magnitude of transfer function associated with minimum phase, unless all the zeros of this function occur at real frequencies ($\lambda = j\omega$). There are problems in which the specified Z_{12} or Y_{12} does not have all its zeros occur at real frequencies, but for which the minimum phase requirement must be met. A possible design procedure in such cases is the following, based upon the so-called *constant-resistance* lattice.

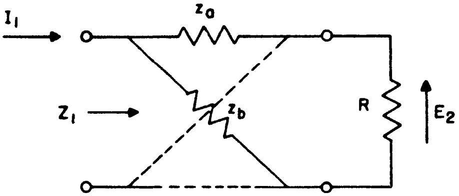


FIG. 6 The symmetrical lattice becomes a so-called constant-resistance network ($Z_1 = R$) if $z_a z_b = R^2$.

With reference to the usual representation of the symmetrical lattice, as shown in Fig. 6, it is a straightforward matter to show that if the pair of impedances z_a and z_b are chosen to fulfil the condition

$$z_a z_b = R^2 \tag{86}$$

then

$$Z_1 = R \tag{87}$$

and

$$Z_{12} = \frac{E_2}{I_1} = \frac{R(R - z_a)}{R + z_a} \tag{88}$$

This network is called a "constant-resistance" lattice.

Normalizing R at one ohm, simplifies the relation 88 to

$$Z_{12} = \frac{1 - z_a}{1 + z_a} \tag{89}$$

or conversely

$$z_a = \frac{1 - Z_{12}}{1 + Z_{12}} \tag{90}$$

If Z_{12} has its poles restricted to the left half-plane (which is, as usual, necessary to insure stability), and if, in addition,

$$|Z_{12}(j\omega)| \leq 1 \tag{91}$$

then the p.r. character of z_a, and hence the realizability of the network, is assured.

The resulting lattice in general is not lossless. To demand a lossless lattice implies $|Z_{12}(j\omega)| \equiv 1$, as may readily be seen from eq. 89 assuming

z_a to be a pure reactance. Such a choice is made if the network is intended to influence phase alone, and it then is obviously not a minimum phase network. In order to obtain a minimum phase network for any other prescribed

$$|Z_{12}(j\omega)|^2 = \frac{G_0 + G_1\omega^2 + \cdots + G_n\omega^{2n}}{H_0 + H_1\omega^2 + \cdots + H_n\omega^{2n}} = \frac{G(\omega^2)}{H(\omega^2)} \leq 1 \qquad (92)$$

it is merely necessary in the formation of

$$Z_{12}(\lambda) = \frac{g(\lambda)}{h(\lambda)} \qquad (93)$$

to construct $g(\lambda)$ from the left half-plane zeros of $G(-\lambda^2)$ just as $h(\lambda)$ must in any case be constructed from the left half-plane zeros of $H(-\lambda^2)$. The polynomial $g(\lambda)$ is now not restricted to be even or odd as in the procedure discussed in Section VIII.

If Z_{12} in eq. 93 is regarded in the factored form

$$Z_{12}(\lambda) = \frac{(\lambda - \lambda_1)(\lambda - \lambda_3) \cdots (\lambda - \lambda_{2n-1})}{(\lambda - \lambda_2)(\lambda - \lambda_4) \cdots (\lambda - \lambda_{2n})} \qquad (94)$$

an important flexibility inherent in the present synthesis procedure becomes evident through recognition that one may readily decompose Z_{12} into the product of components as indicated in

$$Z_{12} = Z_{12}^{(1)} \times Z_{12}^{(2)} \times Z_{12}^{(3)} \times \cdots \qquad (95)$$

By shuffling and reshuffling the frequency factors $(\lambda - \lambda_\nu)$ in the numerator or denominator of eq. 94, one can obtain several distinct decompositions like eq. 95 (observing, however, that where a pair of complex values of λ_ν is involved the entire quadratic factor must, of course, be kept intact). If a lattice is found to correspond to each component $Z_{12}^{(1)}$, $Z_{12}^{(2)}$, etc., their cascade connection evidently realizes Z_{12} because of the constant-resistance character of each component lattice.

It is necessary, of course, to see to it that each component Z_{12}-function fulfils the condition 91, a circumstance that is not assured by the net Z_{12}-function's fulfilling this condition, but one that can always be brought about through inserting an appropriate constant multiplier. Thus the price of being able to decompose an elaborate design into a cascade of simple structures may be a net loss in gain. The larger variety of possible realizations made available through this artifice is in many cases worth the cost.

Still further possibilities may be had through inserting identical arbitrary left half-plane zeros (called *surplus factors*) into numerator and denominator of Z_{12} before the shuffling and partitioning into components is begun.

XII. AN ALTERNATE REALIZATION PROCEDURE FOR TRANSFER FUNCTIONS

When the desired transfer characteristic is to be had from a cascade of vacuum tube amplifier stages, a decomposition of the overall function in the manner indicated in eq. 95 is appropriate in which each component Z_{12}-function represents the transfer function for a single stage. If a pentode tube is used, and if the component Z_{12}-function happens to be p.r. so that it is realizable as a driving point impedance, then the corresponding amplifier stage consists simply of the pentode with the pertinent driving point impedance in its plate circuit. The success of the method depends on the fact that, through the choice of suitable surplus factors, the given overall Z_{12}-function can always be represented as a product of p.r. driving point impedances.*

To appreciate the manner in which such a decomposition of Z_{12} may be accomplished one need merely consider, for example, a typical portion of Z_{12} consisting of the quotient of quadratic factors

$$\frac{\lambda^2 + a\lambda + b}{\lambda^2 + c\lambda + d} \tag{96}$$

which, after multiplication and division by the surplus factor $(\lambda + e)$, can be separated into the product of two factors as follows:

$$\frac{\lambda^2 + a\lambda + b}{\lambda + e} \times \frac{\lambda + e}{\lambda^2 + c\lambda + d} \tag{97}$$

Each of these has a simple network realization as a driving point impedance if $e < a$ and $e < c$, which can always be met (a, b, c, d, e are positive real numbers). The expression 96 may, of course, have a simple driving point realization as it stands; the method involving surplus factors is resorted to only if needed.

XIII. SYNTHESIS OF A LOSSLESS TWO TERMINAL-PAIR NETWORK THROUGH THE LADDER DEVELOPMENT OF z_{22}

This is a method of synthesis that is particularly useful in conjunction with the problem discussed in Section VIII, since it yields the lossless network directly from z_{22} and z_{12} without the necessity of associating an appropriate z_{11}. The network is found through developing z_{22} into a ladder, much as in the Cauer process for realizing a given driving point reactance function, and, after bringing out a terminal pair at the far end of this ladder, regarding it as the desired two terminal-pair network. It must be shown how this process can be carried out so as to produce a network having the desired z_{12}-function.

---

* It is here tacitly assumed that Z_{12} is a minimum phase function.

In this regard it should first be observed that in general all the functions z_{11}, z_{22}, z_{12} associated with a two terminal-pair network have the same poles, since these are physically the natural frequencies of the network with both terminal pairs open-circuited. Degenerate cases may, of course, occur in which this situation is not realized. For example, take any two terminal-pair network and connect in series with one of its terminal-pairs (say at end No. 2) an impedance having poles other than those present in the impedances of the given network. One then has a two terminal-pair network for which z_{22} has these extra poles but z_{11} and z_{12} do not have them. If, however, the desired network is to be developed from a given z_{22}, it is easy to see how such a degenerate case is to be avoided. Namely, either avoid an initial series branch in the development of z_{22}, or if an initial series branch is indicated for other reasons, be sure that the pole or poles it represents are not completely removed from z_{22} (that is, z_{22} minus the impedance of the initial series branch should have the same poles although not with the same residues - as z_{22}). Thus the determination of a network whose z_{12}-function has the same poles as z_{22} presents no difficulties.

It remains to show how the development of z_{22} into a ladder network can be made to yield a z_{12}-function having the proper zeros. So long as the desired zeros occur at real frequencies ($\lambda = j\omega$), an appropriate procedure is easy to find and the resulting network is simple and practical. Although Darlington has shown that the desired procedure can be carried out no matter where the zeros of z_{12} are located in the λ-plane, computations are extremely tedious and the resulting network undesirable (because of close-coupled coils) unless the zeros lie on the j-axis. The following further discussion is, therefore, restricted to this case.

In a two terminal-pair network having a ladder structure it is clear that a zero of the transfer impedance requires either that a series-branch impedance be infinite or that a shunt-branch impedance be zero. Note, however, that an infinite series-branch impedance or a zero shunt-branch impedance does not necessarily cause a zero in the transfer impedance, for the part of the network to the left* of this point, when regarded as a two terminal-pair network in its own right, may have a z_{12}-function with a pole at the same frequency.

The ladder development of a driving point reactance function like z_{22} is ordinarily accomplished through alternately constructing a series branch representing a pole of z_{22} (or of a remainder function) and a shunt branch representing a pole of $1/z_{22}$ (zero of z_{22}, or of a remainder function). At each step the pole in question is *completely* removed from the pertinent driving point function. These poles become the zeros of the transfer function of the two terminal-pair network which the resulting ladder represents.

---

* The source is assumed to be at the left-hand end of the network.

If at a particular step in this procedure (for example, at the construction of a series branch) a pole of the pertinent driving point function were only *partially* removed (by subtracting a pole with smaller residue than that of the driving point function) then this pole does *not* become a zero of the transfer function of the resulting network because the part of the network to the left of this point (as yet undeveloped) has a z_{22}-function and hence a z_{12}-function which still contains this pole. (If the step is the construction of a shunt branch, the same comment applies with reference to admittances.) Thus in the ladder development of a given z_{22}-function, the construction of a particular series or shunt branch may or may not be "zero producing" so far as the resulting transfer function is concerned, depending respectively upon the complete or partial removal of a pole.

It should next be recognized that through the partial removal of a pole of a reactance (resp. susceptance) function, its zeros are shifted; and it is possible through the removal of an appropriate part of an appropriate pole to produce in the subsequent remainder function a zero at *any* stated frequency. In the inverted remainder function this step produces a pole at the same frequency. The subsequent complete removal of this pole produces a zero in the transfer function of the resulting network at that frequency.

The first step, consisting of the partial removal of a pole, may be referred to as a "zero-shifting" step Thus the process of ladder development consists alternately of "zero-shifting" and "zero-producing" steps, continued until all the zeros of the given z_{12}-function have been produced, at which point the development of z_{22} will be completed and the desired network found.

Since the zeros of the given z_{12}-function can be produced in varying sequence, several different networks can in general be found for a given design problem. The variety of possible networks may further be increased through recognizing that one may follow not only that scheme in which series branches represent zero-shifting steps and shunt branches represent zero-producing ones, but also the parallel scheme in which the roles played by series and shunt branches are interchanged; or one may even scramble these two basic procedures.

In problems where the zeros of z_{12} for the desired lossless network do not occur on the j-axis, one may find it expedient to apply a method based upon the following reasoning. The poles of z_{12}, like those of a driving point reactance are all simple and lie on the j-axis, but the residues in these poles, although real, are not necessarily all positive. In fact it is the negativeness of some of these residues that causes the zeros of z_{12} to be located off of the j-axis, for if the residues were all positive, z_{12} would have the same form as a driving point reactance and hence its zeros would surely be on the j-axis.

Suppose z_{12} is expanded into partial fractions and the terms with positive residues grouped to form the function $z_{12}^{(1)}$ while those with negative residues are regarded as forming $-z_{12}^{(2)}$. Thus $z_{12} = z_{12}^{(1)} - z_{12}^{(2)}$ in which $z_{12}^{(1)}$ and $z_{12}^{(2)}$ both have the character of driving point functions and hence have all their zeros on the j-axis.

The relation 63 for Z_{12} is now written

$$Z_{12} = \frac{z_{12}^{(1)}}{1 + z_{22}} - \frac{z_{12}^{(2)}}{1 + z_{22}} \tag{98}$$

Each of these two terms is realizable as a ladder network according to the method discussed above. Joining their output terminals through an ideal transformer (after multiplying each network by an appropriate constant to allow for possible unequal multipliers in the individual z_{12}-functions and to obtain a resultant z_{22} with the multiplier unity) yields a realization for the total function Z_{12}.

Finally, the method of synthesis discussed in this article may be useful in the realization of Z_{12}-functions in cases where the zeros are restricted to the left half-plane because of a minimum phase requirement, and where these zeros do not lie on the j-axis so that a realization in terms of a lossless network terminated in a single load resistance is not possible.

In the expression $Z_{12} = g(\lambda)/h(\lambda)$ both $g(\lambda)$ and $h(\lambda)$ are Hurwitz polynomials. If $g = \mu + \nu$ and $h = m + n$ represent the separation of $g(\lambda)$ and $h(\lambda)$ into their even and odd parts respectively, one can write

$$\begin{aligned} Z_{12} &= \frac{\mu}{m + n} + \frac{\nu}{m + n} \\ &= Z_{12}^{(1)} + Z_{12}^{(2)} \end{aligned} \tag{99}$$

For the realization of each component Z_{12}-function according to the scheme indicated in eqs. 67, 68, and 69, one has

$$Z_{12}^{(1)} = \frac{\mu/n}{1 + (m/n)}; \quad z_{12}^{(1)} = (\mu/n); \quad z_{22}^{(1)} = (m/n) \tag{100}$$

and

$$Z_{12}^{(2)} = \frac{\nu/m}{1 + (n/m)}; \quad z_{12}^{(2)} = (\nu/m); \quad z_{22}^{(2)} = (n/m) \tag{101}$$

Since μ and ν individually have simple zeros restricted to the j-axis (because of the properties of Hurwitz polynomials), the method of ladder development given in this article is applicable to both functions 100 and 101.

After an unbalanced lossless ladder network terminated in its appropriate resistance is found for the realization of each component Z_{12}-function (the terminal resistances for these networks are not necessarily alike since their impedance levels are the means for independently controlling the multipliers in 100 and 101), a back-to-back connection as shown in Fig. 7 may be used to obtain the overall Z_{12}.

Observe that the z_{22}-functions of the individual ladder networks are reciprocal For this reason the two lossless ladders cannot be combined and terminated in a single resistance load as in the process to which eq. 98 applies. It was pointed out earlier that the lossless network terminated in a single resistance cannot realize a minimum phase transfer function with zeros off of the j-axis; the present result is consistent with this statement.

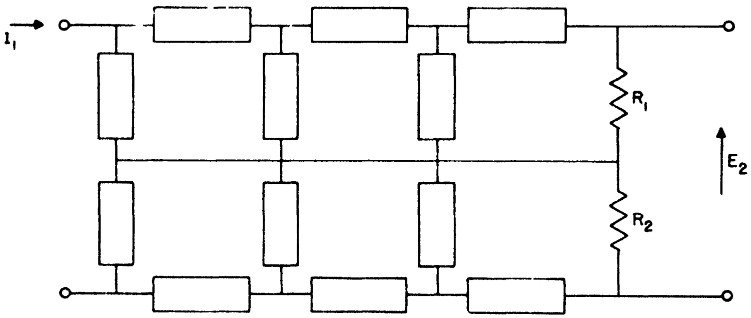


FIG. 7 A form of network in which it is possible to realize any minimum phase transfer function E_2/I_1 The boxes contain lossless elements.

ILLUSTRATIVE EXAMPLES

XIV. BRUNE'S SYNTHESIS PROCEDURE

Let the given driving point impedance be

$$Z(\lambda) = \frac{5\lambda^2 + 3\lambda + 4}{\lambda^2 + 2\lambda + 2} \tag{102}$$

Its real part for $\lambda = j\omega$ is

$$\begin{aligned} \mathrm{Re}[Z(j\omega)] &= \left[\frac{(5\lambda^2 + 4)(\lambda^2 + 2) - 6\lambda^2}{(\lambda^2 + 2)^2 - 4\lambda^2}\right]_{\lambda = j\omega} \\ &= \left(\frac{5\lambda^4 + 8\lambda^2 + 8}{\lambda^4 - 4}\right)_{\lambda = j\omega} = \frac{5\omega^4 - 8\omega^2 + 8}{\omega^4 + 4} \end{aligned} \tag{103}$$

This function has a minimum value equal to unity at $\lambda_1 = j\omega_1 = j1$. Removing a series resistance of 1 ohm leaves (eq. 15)

$$Z_1(\lambda) = Z(\lambda) - 1 = \frac{4\lambda^2 + \lambda + 2}{\lambda^2 + 2\lambda + 2} \tag{104}$$

Next one finds

$$Z_1(\lambda_1) = \frac{-4 + 2 + j1}{-1 + 2 + j2} = \frac{-2 + j1}{1 + j2} = j1 \tag{105}$$

so X in eq. 16 equals +1. This is an example in which the p.r. character of the function is lost if we proceed on an impedance basis. Nevertheless,

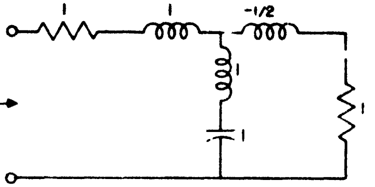


FIG. 8. Realization of the impedance of eq 102 through the Brune procedure.

we continue with $L_1 = X/\omega_1 = 1$ and have

$$Z_1(\lambda) - L_1\lambda = \frac{(\lambda^2 + 1)(-\lambda + 2)}{\lambda^2 + 2\lambda + 2} \tag{106}$$

which is zero at $\lambda = \pm\lambda_1 = \pm j1$ as it should be, but is clearly no longer p.r. The reciprocal of eq. 106, however, has a positive real residue at $\lambda = j$ as may be seen from

$$k = \left[\frac{\lambda^2 + 2\lambda + 2}{(\lambda + j)(-\lambda + 2)}\right]_{\lambda = j} = \frac{1}{2} \tag{107}$$

Removal of the pair of j-axis poles of $(Z_1 - L_1\lambda)^{-1}$ in the Foster manner yields (eq. 17)

$$\frac{\lambda^2 + 2\lambda + 2}{(\lambda^2 + 1)(-\lambda + 2)} - \frac{\lambda}{\lambda^2 + 1} = \frac{2}{-\lambda + 2} = \frac{1}{W(\lambda)} \tag{108}$$

Finally

$$W(\lambda) = -\frac{\lambda}{2} + 1 = L_3\lambda + Z_2(\lambda) \tag{109}$$

Here the remainder $Z_2(\lambda)$ is a resistance of 1 ohm. The network is shown in Fig. 8, in which the parameter values are in henrys, farads, and ohms.

The three inductances may alternately be represented as a pair of closely coupled coils if desired.

XV. DARLINGTON'S PROCEDURE APPLIED TO THE SAME PROBLEM

From $Z(\lambda)$ as given in eq. 102 we compute

$$\begin{aligned}(m_1m_2 - n_1n_2) &= 5\lambda^4 + 8\lambda^2 + 8 \\ &= 5(\lambda^2 + 0.964\lambda + 1.265)(\lambda^2 - 0.964\lambda + 1.265)\end{aligned} \qquad (110)$$

The factor $m_0^2 - n_0^2 = (m_0 + n_0)(m_0 - n_0)$ needed in this case in order to obtain a revised $A(-\lambda^2)$ which is a full square, is the entire expression 110 except for the multiplier 5. From the factored form in 110 it is clear, therefore, that the auxiliary Hurwitz polynomial reads

$$P_0 = \lambda^2 + 0.964\lambda + 1.265 \qquad (111)$$

Augmenting $Z(\lambda)$ in eq. 102 through multiplying and dividing by P_0 yields the revised $Z(\lambda)$

$$Z(\lambda) = \frac{5\lambda^4 + 7.820\lambda^3 + 13.217\lambda^2 + 7.651\lambda + 5.060}{\lambda^4 + 2.964\lambda^3 + 5.193\lambda^2 + 4.458\lambda + 2.530} \qquad (112)$$

Using this function one now finds

$$m_1m_2 - n_1n_2 = 5(\lambda^4 + \tfrac{8}{5}\lambda^2 + \tfrac{8}{5})^2 = A(-\lambda^2) \qquad (113)$$

This function is even in λ^2. Hence the case A formulas apply, and we have from eqs. 42 and 45

$$z_{11} = \frac{m_1}{n_2} = \frac{5\lambda^4 + 13.217\lambda^2 + 5.060}{2.964\lambda^3 + 4.458\lambda} \qquad (114)$$

$$z_{22} = \frac{m_2}{n_2} = \frac{\lambda^4 + 5.193\lambda^2 + 2.530}{2.964\lambda^3 + 4.458\lambda} \qquad (115)$$

$$z_{12} = \frac{\sqrt{5}\,(\lambda^4 + 1.6\lambda^2 + 1.6)}{2.964\lambda^3 + 4.458\lambda} \qquad (116)$$

The partial fraction expansions of these functions read

$$z_{11} = \frac{1}{0.88\lambda} + \frac{0.787\lambda}{\lambda^2 + 1.503} + 1.687\lambda \qquad (117)$$

$$z_{22} = \frac{1}{1.761\lambda} + \frac{0.678\lambda}{\lambda^2 + 1.503} + 0.3375\lambda \qquad (118)$$

$$z_{12} = \frac{1}{1.245\lambda} - \frac{0.730\lambda}{\lambda^2 + 1.503} + 0.754\lambda \qquad (119)$$

from which one observes that the residue condition 51 is fulfilled with the equals sign at each pole.

The synthesis may now follow the pattern of setting down a component lossless two terminal-pair network for each pole and connecting these in series on the input and on the output sides. In this connection one observes that ideal transformers can be avoided through changing the output impedance level by a factor which makes the residues of z_{11}, z_{22}, z_{12} at $\lambda = 0$ alike. This factor is $1.761/0.88 = 2.0$. The revised expressions for z_{22} and z_{12} read

$$z_{22} = \frac{1}{0.88\lambda} + \frac{1.356\lambda}{\lambda^2 + 1.503} + 0.675\lambda \tag{120}$$

$$z_{12} = \frac{1}{0.88\lambda} - \frac{1.032\lambda}{\lambda^2 + 1.503} + 1.066\lambda \tag{121}$$

The resultant network is shown in Fig. 9 in which the two coil pairs are closely coupled, and the one paralleled by capacitance has a negative mutual in order to account for the negative residue of z_{12} in the pertinent pole. Observe that the terminal resistance is 2 ohms because the output impedance level was increased by a factor 2. The resultant driving point impedance is the same as for the network of Fig. 8.

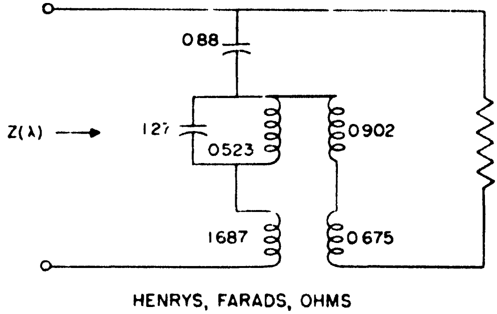


FIG 9 Realization of the impedance of eq 102 through the Darlington procedure (equivalent to network of Fig 8)

XVI. AN ALTERNATIVE METHOD OF SYNTHESIS THAT AVOIDS MUTUAL COUPLING

The success of this procedure (due to R. Bott and R. J. Duffin)[9] depends upon the fact that, if $Z_1(\lambda)$ is p.r. and k is a positive real number, then

$$R(\lambda) = \frac{kZ_1(\lambda) - \lambda Z_1(k)}{kZ_1(k) - \lambda Z_1(\lambda)} \tag{122}$$

is likewise p.r. and has no greater degree of complexity than $Z_1(\lambda)$. Moreover, $R(\lambda)$ may be made to have a pair of conjugate j-axis zeros at the frequency λ_1 where $Z_1(\lambda_1) = jX$ (as in the Brune process) through an appropriate choice for the value of k; whence, solving eq. 122 for $Z_1(\lambda)$

$$Z_1(\lambda) = \frac{Z_1(k)}{(1/R) + (\lambda/k)} + \frac{Z_1(k)}{R + (k/\lambda)} \tag{123}$$

shows that a simplification can be effected through applying Foster's procedure alone. The requirement $R(\lambda_1) = 0$, which yields

$$\frac{Z_1(\lambda_1)}{\lambda_1} = \frac{Z_1(k)}{k} \tag{124}$$

can be met with a positive k only if $X > 0$. However, when $X < 0$ the identical procedure is possible on an admittance basis.

To illustrate with the function 102 for $Z(\lambda)$ one begins as in the Brune process by removing a series resistance equal to the minimum value of Re$[Z(j\omega)]$, leaving $Z_1(\lambda)$ as given in eq. 104. Since $\lambda_1 = j1$ and $Z_1(\lambda_1) = j1$, eq. 124 requires in this example that

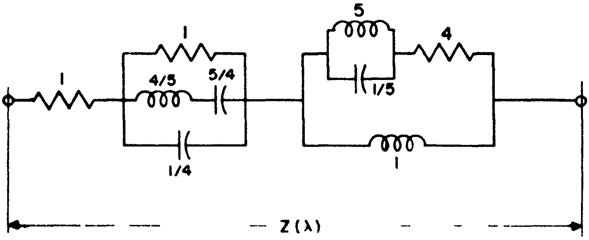


FIG. 10. Realization of the impedance of eq. 102 through the procedure of Bott and Duffin (equivalent to networks of Figs. 8 and 9)

$$Z_1(k) = k$$

or, using eq. 104,

$$4k^2 + k + 2 = k^3 + 2k^2 + 2k$$

The positive real root of this equation is $k = 2$. Thus eq. 122 yields

$$R(\lambda) = \frac{2(\lambda^2 + 1)}{4\lambda^2 + 5\lambda + 4} \tag{125}$$

and $Z_1(\lambda)$ according to eq. 123 is given by

$$Z_1(\lambda) = \frac{2}{\dfrac{4\lambda^2 + 5\lambda + 4}{2(\lambda^2 + 1)} + \dfrac{\lambda}{2}} + \frac{2}{\dfrac{2(\lambda^2 + 1)}{4\lambda^2 + 5\lambda + 4} + \dfrac{2}{\lambda}} \tag{126}$$

Using only the Foster procedure on each of these two terms one finds the network shown in Fig. 10. The left-hand resistance of 1 ohm is the minimum of $Z(j\omega)$ first removed. The other two resistances in the networks realizing the separate terms in eq. 126, are the remainder functions (like $Z_2(\lambda)$ in the Brune process). In more complicated examples, these remainders would be dealt with in the same manner, yielding additional network elements and four subsequent remainder functions which in turn again receive the same treatment. After n cycles one has 2^n network components like the ones shown in Fig. 10. The price of avoiding mutual coupling is in general seen to be a rather large number of network elements.

XVII. Darlington's Procedure Applied to the Synthesis of a Transfer Impedance

Here one begins with a specified $|Z_{12}(j\omega)|^2 = |E_2/I_1|^2$, with reference to the situation shown in Fig. 3. Suppose one is interested in a low-pass prototype filter and has chosen the so-called Butterworth function

$$|Z_{12}(j\omega)|^2 = \frac{1}{1 + \omega^{2n}} \tag{127}$$

For $n = 3$ one then has, according to eq. 55 with $R = 1$,

$$\operatorname{Re}[Z_1(j\omega)] = \frac{1}{1 + \omega^6} = \frac{A(\omega^2)}{B(\omega^2)} \tag{128}$$

in which $Z_1(\lambda)$ is the input impedance as shown in Fig. 3. Here

$$B(-\lambda^2) = 1 - \lambda^6 \tag{129}$$

and the zeros of $B(-\lambda^2)$ are seen to be the sixth roots of unity. The product of left half-plane factors yields

$$m_2 + n_2 = (\lambda + 1)(\lambda^2 + \lambda + 1) = \lambda^3 + 2\lambda^2 + 2\lambda + 1 \tag{130}$$

as the denominator of the input impedance $Z_1(\lambda)$.

In applying the Gewertz method to the determination of the numerator of $Z_1(\lambda)$ according to eqs. 59, one observes that in the present problem $A_0 = 1$ and $A_1 = A_2 = A_3 = 0$, giving

$$\begin{aligned} a_0 &= 1 \\ -2a_0 + 2a_1 - a_2 &= 0 \\ -a_1 + 2a_2 - 2a_3 &= 0 \\ a_3 &= 0 \end{aligned} \tag{131}$$

whence

$$a_0 = 1, \quad a_1 = \tfrac{4}{3}, \quad a_2 = \tfrac{2}{3}, \quad a_3 = 0 \tag{132}$$

The desired input impedance, having the real part 128, therefore is

$$Z_1(\lambda) = \frac{1 + \frac{4}{3}\lambda + \frac{2}{3}\lambda^2}{1 + 2\lambda + 2\lambda^2 + \lambda^3} \tag{133}$$

Synthesis of a network in this example is very simple because of the special form of the real part of Z_1 assumed at the outset (eq. 128). We are dealing here with a real part whose minima are all zero and lie at $\lambda = \infty$, a point on the j-axis. In cases where the minima of the real part of a driving-point impedance are all zero and lie on the imaginary axis, it is clear that the Brune procedure (like Darlington's) will yield a lossless two terminal-pair network terminated in a single resistance, since at the beginning of each cycle one removes a series branch having zero resistance.

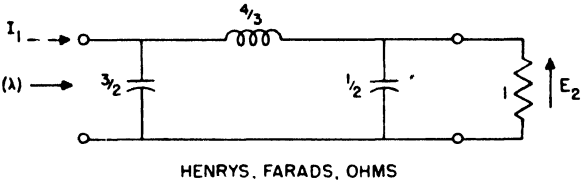


Fig. 11. Network whose transfer impedance is the Butterworth function of eq. 127 for $n = 3$

When all the zeros of the real part lie at $\lambda = 0$ or at $\lambda = \infty$, further simplifications take place, with the result that the synthesis is completed through application of the preamble to the Brune method alone. Thus $Z_1(\lambda)$ in eq. 133 has a zero at $\lambda = \infty$ which can be removed by Foster's method. The inverted remainder again has a zero at $\lambda = \infty$, and so does the inverted remainder after that, etc. The entire process is indicated in the continued fraction development

$$Z_1(\lambda) = \cfrac{1}{\frac{3}{2}\lambda + \cfrac{1}{\frac{4}{3}\lambda + \cfrac{1}{\frac{1}{2}\lambda + \cfrac{1}{1}}}} \tag{134}$$

which possesses the realization shown in Fig. 11.

The function 127 leads in every case to a simple ladder network of this form, with n equal to the total number of reactive elements.

XVIII. Cauer's Method Applied to the Same Problem

Again choosing the function 127 for $n = 3$, one has according to eq. 65

$$|Z_{12}(j\omega)|^2 = \frac{1}{1 + \omega^6} = \left(\frac{g^2}{m^2 - n^2}\right)_{\lambda = j\omega} \tag{135}$$

Here $g(\lambda)$ is simply unity, and $h(\lambda) = m + n$ is found from the left half-plane zeros of $m^2 - n^2$ to be

$$h(\lambda) = \lambda^3 + 2\lambda^2 + 2\lambda + 1 \tag{136}$$

as in the preceding section. Equations 68 then yield

$$z_{22} = \frac{2\lambda^2 + 1}{\lambda^3 + 2\lambda}; \quad z_{12} = \frac{1}{\lambda^3 + 2\lambda} \tag{137}$$

Considering the synthesis procedure discussed in Section XIII, it is significant to note that the zeros of z_{12} all lie at $\lambda = \infty$ (these are the double zeros of $\mathrm{Re}[Z_1(j\omega)]$ dealt with in the previous section.) Since $z_{22} = 0$ at $\lambda = \infty$, no shifting step is needed The development begins by removing the pole at infinity from $1/z_{22}$. The remainder at this stage will be zero at $\lambda = \infty$, and so again it is not necessary to carry out a zero-shifting step. It becomes clear that the synthesis is accomplished, in this case, through the ordinary ladder development of z_{22} as in the Cauer procedure for reactance functions.

The continued fraction expansion

$$z_{22} = \cfrac{1}{\frac{1}{2}\lambda + \cfrac{1}{\frac{4}{3}\lambda + \cfrac{1}{\frac{3}{2}\lambda}}} \tag{138}$$

leading again to the network of Fig. 11, accomplishes the desired end.

XIX. A CONSTANT-RESISTANCE FILTER GROUP

In the example of Section XVII the maximum of $\mathrm{Re}[Z_1(j\omega)]$ is unity. Furthermore, the impedance $Z_1(\lambda)$ found there is minimum reactive. Therefore $1 - Z_1(\lambda) = Z_2(\lambda)$ is surely p.r. This complementary impedance is found from eq. 133 to be

$$Z_2(\lambda) = \frac{\frac{2}{3}\lambda + \frac{4}{3}\lambda^2 + \lambda^3}{1 + 2\lambda + 2\lambda^2 + \lambda^3} \tag{139}$$

It is interesting to note that $Z_2(\lambda) = Z_1(1/\lambda)$, a situation that always results with the Butterworth function 127 because 1 − (this function) is the same as this function with ω replaced by $1/\omega$.

The network realizing eq. 139 may be found through a development of $Z_2(\lambda)$ similar to that of $Z_1(\lambda)$ given by eq. 134, or directly from Fig. 11 observing that $\lambda \rightarrow 1/\lambda$ implies the replacement of an inductance by a capacitance of reciprocal value, and vice versa. The result is shown in

Fig. 12. The associated Z_{12}-function is complementary to that relating to the network of Fig. 11. The two networks placed in series at their input terminals constitute a low-pass, high-pass, constant-resistance filter group. So do the reciprocal networks placed in parallel at their inputs.

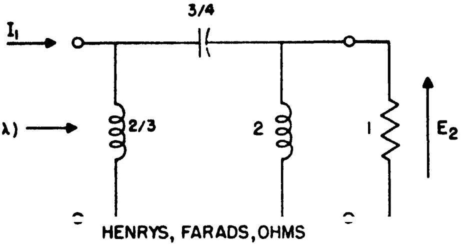


FIG. 12. Network complementary to that of Fig. 11; the sum of the driving point impedances of these networks is constant ($Z_1 + Z_2 = 1$), as is the sum of the squared magnitudes of their transfer impedances

XX. The Same Transfer Function Realized through a Lossless Network with Resistance Loading at Both Ends

Here the function t^2 of eq. 76 is set equal to the Butterworth function; that is, for $n = 3$

$$t^2 = \frac{1}{1 + \omega^6} \tag{140}$$

and according to eq. 78

$$\rho^2 = \frac{\omega^6}{1 + \omega^6} \tag{141}$$

Constructing the numerator and denominator polynomials of $\rho(\lambda)$ in the usual manner gives

$$\rho(\lambda) = \frac{\lambda^3}{1 + 2\lambda + 2\lambda^2 + \lambda^3} \tag{142}$$

and the input impedance to the lossless network, eq. 77 for $R_1 = 1$, becomes

$$Z_1(\lambda) = \frac{1 + 2\lambda + 2\lambda^2}{1 + 2\lambda + 2\lambda^2 + 2\lambda^3} \tag{143}$$

The process of realization follows the same pattern as in the example of Section XVII. Thus the continued fraction development

$$Z_1(\lambda) = \cfrac{1}{\lambda + \cfrac{1}{2\lambda + \cfrac{1}{\lambda + \cfrac{1}{1}}}} \tag{144}$$

leads to the network of Fig. 13.

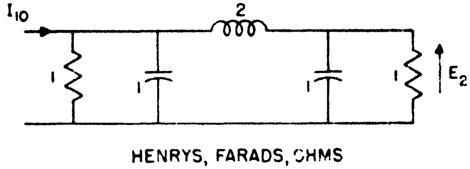


FIG. 13 Lossless coupling network with resistive loading at both ends, having the same transfer impedance as that in Fig 11.

XXI. Realization through a Cascade of Amplifier Stages

As shown in Article XVIII, the Butterworth function 127 for $|Z_{12}(j\omega)|^2$ (with $n = 3$) yields

$$Z_{12}(\lambda) = \frac{1}{1 + 2\lambda + 2\lambda^2 + \lambda^3} = \frac{1}{(\lambda + 1)(\lambda^2 + \lambda + 1)} \tag{145}$$

which may be represented as the product of three p.r. driving point functions, as follows

$$Z_{12}(\lambda) = \frac{\lambda + 1}{\lambda^2 + \lambda + 1} \times \frac{1}{\lambda + 1} \times \frac{1}{\lambda + 1} \tag{146}$$

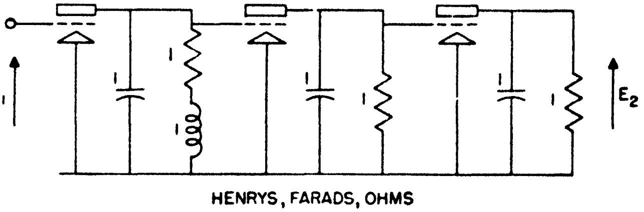


FIG. 14. Cascade of pentode amplifier stages whose transfer function E_2/E_1 is equal (except for a constant multiplier) to the Butterworth function realized through the networks of Figs. 11 and 13.

Thus the circuit of Fig. 14, in which the tubes are assumed to be pentodes, yields another form of realization for this same transfer function. This form of realization (for obvious reasons) will always have minimum-phase properties.

XXII. FURTHER ILLUSTRATION OF THE LADDER DEVELOPMENT PROCEDURE

Suppose the pair of impedance functions

$$z_{11} = \frac{16\lambda^4 + 9\lambda^2 + 1}{26\lambda^5 + 21\lambda^3 + 3\lambda}, \quad z_{12} = \frac{(4\lambda^2 + 1)(\lambda^2 + 1)}{26\lambda^5 + 21\lambda^3 + 3\lambda} \tag{147}$$

are given, and a corresponding lossless two terminal-pair network is to be found through a ladder development of z_{11} carried out in such a way as to produce the zeros called for in z_{12}. The latter occur at $\lambda^2 = -1$, $\lambda^2 = -\frac{1}{4}$, $\lambda = \infty$. In the ladder development of z_{11}, let the zeros of z_{12} be produced in this order.

The object of the first step is to remove a branch so as to yield a remainder with zeros at $\lambda = \pm j$. We begin by computing

$$z_{11}(j1) = \frac{1 - 9 + 16}{j(3 - 21 + 26)} = \frac{1}{j} = -j \tag{148}$$

Since this reactive value is negative (indicating that a negative inductance should be removed), we decide to shift to an admittance basis and remove a positive capacitance $c_1 = 1$ farad from $1/z_{11} = y_1$, leaving

$$y_1 - c_1\lambda = \frac{2\lambda(\lambda^2 + 1)(5\lambda^2 + 1)}{16\lambda^4 + 9\lambda^2 + 1} = y_2 = \frac{1}{z_2} \tag{149}$$

The impedance z_2 thus has a pole at $\lambda = \pm j$, with the residue

$$\left[\frac{16\lambda^4 + 9\lambda^2 + 1}{2\lambda(\lambda + j)(5\lambda^2 + 1)}\right]_{\lambda=j} = \frac{1}{2} \tag{150}$$

According to Foster's procedure we remove a series branch with the impedance $\lambda/(\lambda^2 + 1)$, leaving

$$z_3 = z_2 - \frac{\lambda}{\lambda^2 + 1} = \frac{6\lambda^2 + 1}{2\lambda(5\lambda^2 + 1)} = \frac{1}{y_3} \tag{151}$$

Next we must produce a zero at $\lambda = \pm j\frac{1}{2}$, so we compute

$$y_3(j\tfrac{1}{2}) = \frac{j(-\frac{5}{4} + 1)}{-\frac{6}{4} + 1} = j\tfrac{1}{2}, \tag{152}$$

whence the following step is to remove a shunt capacitance $c_3 = y_3(j\frac{1}{2})/(j\frac{1}{2}) = 1$ farad and have left

$$y_3 - c_3\lambda = \frac{\lambda(4\lambda^2 + 1)}{6\lambda^2 + 1} = \frac{1}{z_4} \tag{153}$$

Thus z_4 has the required poles at $\lambda = \pm j\frac{1}{2}$ with residue

$$\left[\frac{6\lambda^2 + 1}{4\lambda(\lambda + j\frac{1}{2})}\right]_{\lambda=j\frac{1}{2}} = \frac{1}{4} \tag{154}$$

We remove this pole completely through taking out a series branch with the impedance $2\lambda/(4\lambda^2 + 1)$, leaving

$$z_5 = z_4 - \frac{2\lambda}{4\lambda^2 + 1} = \frac{1}{\lambda} \tag{155}$$

This remainder represents a final shunt capacitance of 1 farad, which produces the remaining zero of z_{12} at $\lambda = \infty$. The resulting network is shown in Fig. 15.

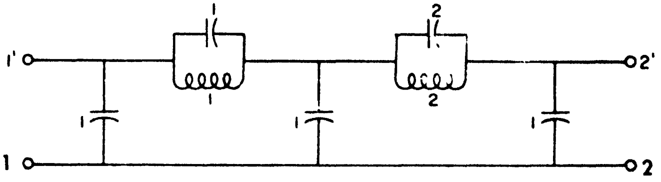


FIG 15 Lossless two terminal-pair network having the driving point and transfer impedances of eq 147.

An alternative network is obtained through taking the zeros of z_{12} in a different order. Suppose we choose the order: $\lambda = \infty$, $\lambda^2 = -1$, $\lambda^2 = -\frac{1}{4}$. As before, the procedure begins with the removal of a shunt capacitance, but this time we want the step to be zero producing, not merely zero shifting. Hence from $1/z_{11}$, we remove the pole at $\lambda = \infty$ completely. The residue in this pole is $^{26}/_{16} = 1.625$; hence the first shunt capacitance becomes $c_1 = 1.625$ farads, and the inverted remainder is found to be

$$z_1 = \frac{16\lambda^4 + 9\lambda^2 + 1}{6.375\lambda^3 + 1.375\lambda} \tag{156}$$

Now we must shift a zero to the points $\lambda = \pm j$, so we compute

$$z_1(j1) = \frac{16 - 9 + 1}{j(-6.375 + 1.375)} = j1.6 \tag{157}$$

which indicates that one should next remove a series inductance $l_1 = z_1(j1)/j1 = 1.6$ henrys, leaving

$$z_1 - l_1\lambda = \frac{(5.8\lambda^2 + 1)(\lambda^2 + 1)}{6.375\lambda^3 + 1.375\lambda} = z_2 \tag{158}$$

The residue of $1/z_2 = y_2$ at $\lambda = j1$ is

$$\left[\frac{6.375\lambda^3 + 1.375\lambda}{(5.8\lambda^2 + 1)(\lambda + j)}\right]_{\lambda=j} = \frac{25}{48} \tag{159}$$

so the removal of this pair of poles reads

$$y_2 - \frac{(\frac{25}{24})\lambda}{\lambda^2 + 1} = y_3 = \frac{(\frac{1}{3})\lambda}{5.8\lambda^2 + 1} = \frac{1}{z_3} \tag{160}$$

Next we compute

$$z_3(j\tfrac{1}{2}) = j2.7 \tag{161}$$

and hence remove a series inductance $l_3 = j2.7/j\tfrac{1}{2} = 5.4$ henrys, leaving

$$z_3 - l_3\lambda = \frac{4\lambda^2 + 1}{(\frac{1}{3})\lambda} = z_4 = \frac{1}{y_4} \tag{162}$$

This y_4 represents a final shunt branch, producing the desired zeros of z_{12} at $\lambda = \pm j\tfrac{1}{2}$. The network for this development is shown in Fig. 16.

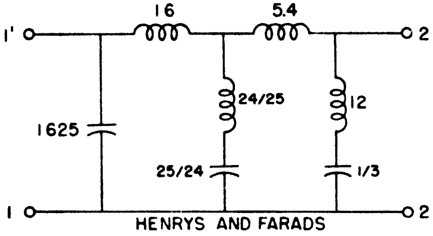


FIG. 16. Lossless two terminal-pair network equivalent (with respect to z_{11} and z_{12}) to that in Fig. 15.

Although the networks of Figs. 15 and 16 have the same z_{11}-function, and the same z_{12}-function except possibly for a difference in constant multipliers, it should be noted that their z_{22}-functions are not the same. Thus the z_{22} of Fig. 15 has only the poles of z_{11} and z_{12} (and the residue condition is consistently fulfilled with the equals sign), while in the network of Fig. 16, z_{22} has in addition a pole at $\lambda = \infty$, which is brought about because the zero-shifting steps here are represented by series inductive branches.

It is not always possible to carry out the desired ladder development so that the residue condition among z_{11}, z_{22}, z_{12} becomes fulfilled with the equals sign (because of the way in which the zero-shifting steps must be carried out so as to avoid negative elements) unless one is willing to accept mutual coupling, perhaps even close coupling or an ideal transformer.

REFERENCES

1. Foster, R. M. *Bell System Tech. J.*, **3**, 259 (1924).
2. Cauer, W. *Arch. Elektrotech.*, **17**, 355 (1927).
3. Brune, O. *J. Math. Phys.*, **X**, 3, 191 (1931).
4. Darlington, S. *J. Math. Phys.*, **XVIII**, 4, 280 (1939).
5. Gewertz, C. M. Network Synthesis. Williams and Wilkins, Baltimore, 1933, pp. 142–149.
6. Bode, H. W. Network Analysis and Feedback Amplifier Design. D. Van Nostrand, New York, 1945, pp. 203–206.
7. Cauer, W. *E.N.T.*, **16**, 6, 161–163 (1939).
8. Darlington, S. *J. Math. Phys.*, **XVIII**, 4, 269 (1939).
9. Bott, R., and Duffin, R. J. *J. Applied Phys.*, **20**, 8, 816 (1949).

Communication Theory

MEYER LEIFER AND WILLIAM F. SCHREIBER

Sylvania Electric Products Inc., Bayside, New York

CONTENTS

Page
I. Introduction 306
II. The Development of the Theory 307
  1. Hartley 307
  2. Gabor 309
  3. Tuller 311
  4. Goldman 313
  5. Coded Transmissions 314
  6. The Maximized Information Function 316
  7. Bandwidth and Time 317
  8. Wiener 318
III. The Synthesis of the Theory 320
  1. Shannon 321
  2. A Geometrical Approach to the Modified Hartley Law 322
  3. Reformulation of the Problem 323
  4. The Measure of Information 324
  5. Coding for the Discrete Source and Noiseless Channel 329
    a. Entropy and the Ordering of Messages 329
    b. The Fundamental Theorem 331
    c. The Coding Process 332
    d. The Coding Delay 333
  6. Coding for the Discrete Source and Noisy Channel 334
    a. The Effect of Noise 334
    b. The Fundamental Theorem 335
  7. The Capacity of a Communication Channel 336
    a. The Discrete Noiseless Channel 336
    b. The Discrete Channel with Noise 336
    c. The Continuous Source and Noisy Channel 337
IV. Applications to Television 339
  References 343

SUMMARY

This paper summarizes and critically examines the work of the principal contributors to basic communication theory. These developments provide a new way of thinking about transmission systems and furnish tools useful in evaluating their performance. In particular, this theory

provides a definition and a measure of the quantity of information in a message. It also provides a measure of the capacity of a communication channel to transmit information. These definitions and measures are shown to be common to all modulation systems and, therefore, provide means for measuring their relative efficiency and for explaining such phenomena as noise improvement in wide-band systems.

In several of the basic problems of general communication theory, attention will be focused upon the concepts rather than the mathematical details. It will not be feasible to review completely the beautiful mathematical structure that Shannon has constructed, nor to do more than select those portions of the work dealing most directly with the basic concepts. An understanding of the assumptions and the principles forms the background from which new problems are attacked and existing systems are evaluated; a way of thinking about communication problems is provided which is at once logical and rigorous. These concepts are developed in terms of the statistics of messages, signals and interference.

I. INTRODUCTION

Communication is one of man's most widespread activities and is essential to his way of life. For this reason he has devoted a substantial portion of his effort to developing and improving means of communication. As a result, there is such a variety of communication systems that it is a difficult task to catalogue them and an even harder one to make up a definition applicable to all of them.

One usually thinks of communication as a transfer of information from a source to a destination. However, in its most general sense, communication exists whenever the state of one entity (the destination) is influenced by the state of another entity (the source). The reason for such an inclusive definition is that, while investigating the fundamental phenomena of such obvious communication systems as voice, music, and television broadcasting, principles have been developed which are directly applicable to other systems where the information transfer process is not so obvious. Such systems include servo-mechanisms, control devices, radar, radio navigation systems, and computers.

The primary purpose of this paper is to present concepts of information and informational capacity especially applicable to electrical communication with a view to developing criteria of excellence by which the performance of all electrical communication systems can be measured. By way of illustration, these criteria will be applied to an analysis of television picture transmission, and on this basis possible methods for reducing frequency bandwidth will be discussed.

II. THE DEVELOPMENT OF THE THEORY

1. Hartley

The most significant early contribution to the theory is found in a paper presented at the International Congress of Telegraphy and Telephony in 1927 by R. V. L. Hartley.[1] A quantitative measure of the information in a message was developed, applicable to both code and voice transmissions, and based on it, a measure of the capacity of a communication channel in terms of its bandwidth was derived. Further research has altered these results only to the extent of clarifying the concept of a "symbol" when noise is present, and removing from the theory the implicit limitations of amplitude modulation.

Hartley considered a message to be a sequence of choices of possible symbols. In Morse code the symbols are the dot and dash. In teletype code the symbols are on and off. In a continuous signal, such as music, the symbols are the successive amplitudes of the signal wave. If there are s different symbols available at each of n successive independent choices, the number of different sequences that might be selected is s^n. In teletype with 2 symbols, on and off, and 5 choices per code group, there are $2^5 = 32$ different code groups.

A measurable characteristic of a message is that it is one selected from a certain number of possible messages. The larger this number of possible messages, the more information is conveyed by one. Therefore the number of possible messages is a measure of the quantity of information in a single message. For example, a page of text in Basic English has less meaning than a similar quantity of standard English because the latter, having more words, has a more exact meaning for each, and hence can express a meaning more precisely in a given number of words. Hartley chose not to use this measure directly, however, because he also desired that the information measure be proportional to the length of the message. This also is in accordance with experience, since we expect two pages to contain twice as much information as one. Since the length of a message is proportional to the number n of characters in it, we have

$$H = \text{information} = Kn$$

where K is a proportionality constant.

For two messages of different length, constructed from a different number of symbols, but both selected from an equal number of possible messages, it is convenient to describe them as containing the same quantity of information.

Let

$$s_1^{n_1} = s_2^{n_2}$$
$$H_1 = K_1 n_1 = H_2 = K_2 n_2$$

Therefore,

$$\frac{K_1}{K_2} = \frac{n_2}{n_1} = \frac{\log s_1}{\log s_2}$$

and K must have the form

$$K = K_0 \log s$$

giving

$$H = K_0 n \log s$$

K_0 may be absorbed in the base of the logarithm, hence

$$H = n \log s = \log s^n = \log M$$

where M is the number of possible messages.

It is evident, therefore, that in order to make the measure of information content proportional to the length of the message, and also a function of the number of possible messages, the appropriate measure is the logarithm of the number of possible messages.

Having set up a suitable measure of information in a message, Hartley then attempted to find how rapidly message symbols might be sent through a channel of limited bandwidth. Assuming that s, the number of available symbols, is fixed by the character of the source of information or by other factors, it is necessary to determine n the number of symbols, which can be sent in a period of time equal to the message duration; in other words, the *rate* at which symbols can be transmitted.

A restricted frequency range or bandwidth in a circuit implies the presence of reactive elements, either inductive or capacitive. This in turn implies the storage of energy in magnetic or electric fields when a message symbol is received Since the stored energy can be dissipated only at a finite rate in any physical network, there is a minimum allowable time between message symbols. This minimum time is the time necessary for the energy stored by one symbol to be sufficiently dissipated so that the next symbol following may be resolved.

In order to find out how the rate at which symbols can be transmitted depends upon the network bandwidth, Hartley analyzed the effect on the intersymbol interference of changing the bandwidth by varying the reactance of the elements. In general, we can state that any transformation in a network which causes its response to occur at time t k instead of time t, and therefore causes the disturbance following a symbol to be damped out k times as fast, will permit the symbol rate to be multiplied by k with the same interference. Since the response of a linear network can be expressed as a sum of terms such as

$$i = Ae^{-\alpha t} \cos (\omega t - \theta)$$

it is easily seen that multiplying α, the damping constant, and ω, the resonant frequency, by k, this result is achieved. From filter theory, this corresponds to dividing the reactance of the elements by k and leaving their resistive parts unchanged, which means multiplying the bandwidth by k. Therefore, in so far as the effect of limited bandwidth may be represented by an energy-storing electrical network, the symbol rate and, therefore, the rate of transmission of information, is proportional to the bandwidth

In recent years it has become evident that certain features of Hartley's work require extension in order to permit adequate evaluation of all factors. The definition of quantity of information in a message depends on the message being constructed by choosing from a finite number of discrete symbols where the message consists of a definite number of such selections. In the case of speech or any other continuous message function, an exact specification of the message would involve an infinite number of levels or symbols from which to choose. This would seem to imply that an infinite quantity of information is contained in a continuous message function. However, this is contrary to common experience. It is evident that one or more factors exist which limit recognition to a finite number of levels or symbols.

A further problem arises in connection with Hartley's discussion of the informational capacity of a given frequency band. The model he employed to represent a transmission channel of limited bandwidth is an electrical circuit, and it is demonstrated that multiplying the bandwidth of this particular circuit by k permits sending symbols k times as fast with the same relative interference. Consequently, the demonstration applies only to the relative allowable symbol rate of two circuits of the same configuration but with different bandwidths. It does not apply to circuits having different configurations, nor does it apply to any channel of limited bandwidth, since it cannot be represented as an electrical network. A final difficulty in the analysis is that it is only in amplitude modulation that the information rate is equal to the signal symbol rate. In more complicated systems, especially coded modulation, a signal symbol may contain either more or less information than was contained in the original message symbol.

2. Gabor

An interesting extension of Hartley's analysis was made by D. Gabor[2] who proceeded to place the theory on a quantitative basis and develop the essential uncertainties in the concepts of bandwidth and transmission time which are basic in communication. Unfortunately, as will be discussed below, he implicitly neglected Hartley's definition of information, thereby restricting the utility of his results to noncoded transmissions.

Gabor assumed that the message to be transmitted may be represented as a continuous function of time limited to a period T. Such a function can be represented within T by a Fourier series in which the terms are separated in frequency by $1/T$, the terms extending to infinite frequency. The message may be represented by the coefficients of the series instead of the original function of time. In any case, if the message is now confined to the bandwidth $(f_2 - f_1)$, there will be $(f_2 - f_1)T$ sine terms and $(f_2 - f_1)T$ cosine terms, or $2(f_2 - f_1)T$ coefficients in all. Hence, a message limited to a period T and a bandwidth W is fully represented by just $2TW$ numbers. Such a message can, therefore, convey the data represented by $2TW$ numbers *and no more*. This argument satisfies some of the objections previously made to Hartley's principle because it no longer depends upon the configuration of a particular network and it removes from the theory the unprecise concept of intersymbol interference.

It is implied in Gabor's work that all messages consisting of the same number of symbols contain the same amount of information. This, as in Hartley's work, implies only certain kinds of modulation, because messages of the type considered are of the form, say, of successive amplitudes of an amplitude-modulated speech wave only. In the general case, however, the signal to be transmitted may have a much more complicated relationship to the original message than the output of a microphone has to the sound waves reaching it. For example, if the successive speech amplitudes are represented by numbers, it would be possible to send all these numbers as a single voltage level whose amplitude in some measurable units gives the same succession of numbers.

It will be evident, as further developments are discussed, that the utility of Gabor's analysis remains essentially limited because of the preceding implicit assumption and the neglect of the effects of noise in the communication system. However, we are indebted to Gabor for an extended treatment of the basic uncertainties in the meaning of bandwidth and time as used in these discussions. Inasmuch as communication signals have both time duration and frequency spectral properties, they may be plotted as areas upon a frequency-time plane. If the time duration of a signal is exactly specified, then the area is sharply defined along the time axis, but is unbounded along the frequency axis. Similarly, if the frequency extent of a signal is exactly specified, the time duration is uncertain. Upon these properties, so similar to the uncertainty relations existing in quantum mechanics, Gabor has constructed a similar mathematical approach. Effective duration and effective frequency width of a signal are defined by the rms deviations of the signal from the mean epoch and the mean frequency. Aside from these statements, the further development represents an interesting digression from the principal channel of this paper.

3. Tuller

In the discussion of Hartley's work, it was shown that the information conveyed by a single symbol or datum is proportional to the logarithm of the number of different values it may have. Hartley did not carry this idea over into his analysis of channel capacity, apparently because at that stage of development the connection between random disturbance in the channel and number of possible symbols was not known. The problem, then, of evaluating the information contained in a message of T seconds duration in a bandwidth of W cycles per second, or the capacity of a communication channel W cycles per second in width available for a period T, reduces to one of finding the number of different values which may be assumed by the signal. This was demonstrated by Tuller,[3] who pointed out that, using Hartley's criterion, an infinitely large amount of information would be contained in a signal which could be measured with an infinite degree of precision. The transmission could therefore take place at an infinite rate through a channel if an infinite number of inputs could be correctly recognized at the output. The output signal need not be identical to the input signal, however, for if the channel alters the signal in some known fashion, precise knowledge of this change will permit recognition of the original signal. If, on the other hand, the signal is changed in some unknown or random fashion the number of input signals which may be recognized and the capacity of the channel become finite.

The finite capacity of the channel may be expressed in terms of the relative magnitudes of the signals and of the fluctuations which are imposed upon them during transmission. Tuller postulated that if the fluctuation is of the nature of random noise a change in signal amplitude can be recognized only if it is at least equal to the rms value N of the noise. If S is the rms value of the signal, just $(1 + S/N)$ different signal symbols can reliably be recognized. Furthermore, these symbols may occur at the rate of only $2W$ per second, because that quantity describes a signal confined to the bandwidth W.

Hartley's expression for the information H contained in a message containing n symbols, where s different symbols are available, may be written

$$H = \log s^n = n \log s$$

where $n = 2TW$
$s = 1 + S/N$
as given respectively by Gabor and Tuller. Then

$$H = 2TW \log (1 + S/N)$$

which is the modified or new Hartley law. It is thus seen that the capacity of a channel will increase if the signal-to-noise ratio is increased. Further, the law implies that neither bandwidth, signal-to-noise ratio, nor time alone, inherently limits the channel capacity. A given message, originally transmitted at a certain rate, occupying a certain bandwidth, and having a specified received signal-to-noise ratio, can be transmitted with any two of these factors reduced by any desired amount, provided the third factor is correspondingly increased. The form in which the modified Hartley law is stated must not be taken in such a way as to lead to the conclusion that these factors, time, bandwidth, signal-to-noise ratio, can be varied at will, as the degree to which the magnitudes may be varied depends upon the modulation system employed. In amplitude modulation, for example, the channel bandwidth should be equal to the message bandwidth. For a given message, a greater channel bandwidth does not improve the transmission and does not permit use of a lower signal-to-noise ratio. A channel bandwidth narrower than the message bandwidth will inevitably result in loss of part of the information, and this loss cannot be prevented by raising the signal-to-noise ratio.

A portion of Tuller's paper describes a system capable of infinite information transfer rate in a noiseless channel, as indicated in the modified Hartley law. To support this contention he uses the idea of a low-pass filter to represent the channel and states that the filter transfer characteristic completely determines its output for a given input. By making measurements of the output, the input signal can be reconstructed and hence, for any single input pulse, the output will be known at any subsequent time. It is then possible to subtract from the channel output that part due to left-over energy of previous pulses or signals, and so recognize each output separately.

There are several weaknesses in an analytical scheme based upon ideal filter theory. First, in speaking of a channel of limited bandwidth, one implies that adjacent channels may be used for other transmissions without interference. The attenuation of the filter would have to be infinite outside the pass band, since noiseless transmission is assumed. According to the Paley-Wiener[1] criterion, such a filter is not even a theoretical possibility. Tuller recognizes this in another part of the paper and states that he uses such a filter only as a first approximation. However, in dealing with noiseless systems, this approximation does not suffice.

If the required ideal filter were realizable, and if it is required to recover the input exactly, then there must be a one-to-one correspondence in waveform between input and output. If the input contains components outside the passband, these components will not appear in the output, hence no one-to-one correspondence will exist. Therefore, a knowledge of the output will not permit complete recognition of the input.

In the example under discussion, the form of the pulses is invariant so that the spectral components outside the passband are not independent of the components within the passband. Hence, the latter together with the filter characteristics are sufficient for a complete specification of the signal at the receiver.

It must be emphasized, however, that the magnitudes of the transmission factors are not necessarily interchangeable in any modulation system. In general, only coded systems exhibit this flexibility. The new Hartley law establishes theoretical limits of rate of transmission which may be realized only by a particular type of modulation.

4. Goldman

A valuable contribution to the theory of communication was made by Stanford Goldman[9] whose work consisted principally in providing an intuitive explanation of some of the physical aspects of the modified Hartley law. He pointed out that a signal limited to a bandwidth W and duration T can be described by various sets of $2TW$ numbers. These may be, as in Tuller's paper, the amplitudes of the signal every $1/2W$ seconds, or they may be the coefficients of the sine and cosine terms of a Fourier series. They may also be the TW amplitudes and TW phase angles of a Fourier series.

Goldman went beyond Tuller in putting the recognition of signals in the presence of noise on a probability basis. He defined the noise level of a signal as the probability that, for a given average noise power, the received signal is not a message but merely a fluctuation. By using the characteristics of random noise, he showed that this probability depends only on the average signal power and not on its distribution in time or frequency.

Goldman's work in this respect does not apply directly to ordinary communication systems, because here one is concerned with a given quantity of information on the average received in spite of the noise—*not* with the case where the entire signal is obscured.

A valuable feature of this work is the physical explanation of noise improvement in wide-band systems. The new Hartley law indicates that it is possible to transmit messages at lower signal-to-noise ratios by increasing the bandwidth. The mechanism by which this is accomplished is the introduction of a certain coherence, or predictable relationship among various parts of the signal, which amounts to a redundancy, either in the frequency domain or the time domain.

Another phenomenon mentioned by Goldman is that of the threshold effect in wide-band systems. If the modified Hartley law is examined, it is noticed that the quantity of information, considered as a function of the signal-to-noise ratio, changes most rapidly for small values of signal-to-noise ratio. This means that the maximum information that can be transmitted through a channel of a given bandwidth decreases very rapidly below a certain value of signal-to-noise ratio. If one considers the mechanism described above for noise improvement in wide-band systems, it becomes apparent why this happens. Each separate sideband operates at a noise level higher than the overall figure. The errors of the spectral components are corrected by comparison with each other, and the chance that all are wrong (perturbed) in the same manner at the same time is small.

As the noise level rises the various sidebands become more and more perturbed, and their relationship in phase and frequency, which was established by modulation in the transmitter, becomes less evident. Eventually, the level of coherence becomes too low for the detector to distinguish, and the noise discrimination property becomes lost.

Goldman explains this phenomenon somewhat mathematically in the following manner. If the signal-to-noise level of each sideband establishes a probability p ($p < 1$) that the signal is merely noise, then r repetitions in r sidebands reduce this probability to p^r. It is readily seen that as p rises, p^r rises very rapidly, and the total output becomes very noisy.

5. *Coded Transmissions*

Tuller developed two other concepts of great importance in information theory: the concept of coded transmission and the concept of the maximized information function.

If one examines the various methods of modulation, it is found that there are two general classes. In the uncoded systems one symbol in the message is transformed into one symbol in the signal. For example, in amplitude modulation each possible amplitude of the intelligence to be transmitted results in a particular amplitude of radio-frequency signal, while in pulse-time modulation each possible amplitude of intelligence results in a particular pulse position.

There is another class of modulation systems, the coded systems, of which pulse-code modulation is the best-known example. In the coded systems each message symbol or amplitude is transformed into a number of signal symbols. Systems also exist where several message symbols are transformed into one signal symbol.

Tuller shows, in a general way, that only with coded systems is one able to achieve full exchange of bandwidth for signal-to-noise ratio. The modified Hartley law,

$$H = 2WT \log (1 + S/N)$$

gives the information capacity of any channel of bandwidth W, signal-to-noise ratio S/N, and duration T. In a communication system one commonly refers to two such channels: the primary information channel—before modulation or after detection and the transmission channel. If W, S, and N refer to the former and W', C (for carrier), and N' to the latter, one can write for an ideal case,

$$2WT \log (1 + S/N) = 2W'T \log (1 + C/N')$$

from which

$$(1 + S/N) = (1 + C/N')^{W'/W}$$

Here W'/W is the bandwidth expansion factor, and the second expression shows how the S/N ratio can be increased, at least ideally, by increasing the bandwidth. One can now investigate the extent to which various systems of modulation can attain this improvement.

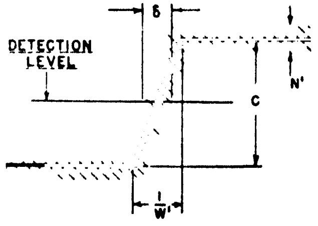


FIG. 1. Uncertainty in pulse position due to noise.

If the bandwidth in an uncoded wide-band system is doubled, the transmission symbol can be located twice as accurately. This will be illustrated for pulse-time modulation. In Fig. 1, δ is the time uncertainty in the position of the pulse due to noise. Since the rise time of the pulse is about $1/W'$, the time uncertainty $\delta = 1/W'(C/N')$. For an rms time uncertainty δ and a maximum pulse excursion L, there are L/δ discrete positions of the pulse, or symbols, which is the same as saying that the signal-to-noise ratio $S/N = L/\delta$. Hence

$$S/N = LW'(C/N')$$

which means that the signal-to-noise ratio is proportional to the bandwidth and hence proportional to the bandwidth expansion. This is not nearly as large an increase as the Hartley law states is possible, and it can generally be shown by similar argument that uncoded systems like pulse-time modulation are inherently incapable of achieving efficient exchange of bandwidth and signal-to-noise ratio.

Coded systems, on the other hand, can be made to take full advantage of added bandwidth. For example, if the bandwidth is doubled, twice as many signal symbols can be transmitted in a given time. If each message symbol is now represented by two signal symbols, and each one has $(1 + C'/N')$ possible values, the combination of two symbols has $(1 + C'/N')^2$ possible values, which is the same as saying that

$$(1 + S/N) = (1 + C'/N')^2$$

where the exponent is the bandwidth expansion factor. In general,

$$(1 + S/N) = (1 + C'/N')^{W'/W}$$

is the relationship governing the noise improvement which the Hartley law states is possible.

In summary, the two examples chosen yield a linear increase in signal-to-noise ratio with an uncoded system, and an exponential increase for a coded system as the bandwidth is increased. This is the real meaning of the Hartley law

6. The Maximized Information Function

The Hartley law states the maximum amount of information which can be transmitted through a transmission link, but not how much information actually is contained within a particular message which is transmitted. It has been shown that the significant point about a particular message is that it is selected from a set of possible messages. There are many cases where all messages are not equally probable. In English, for example, all the letters do not occur with equal frequency. In fact, for messages of great enough length, one can ordinarily predict the letter frequency. Furthermore, one can also predict the frequency with which one letter will follow any other letter; for example, *u* always follows *q*, *h* frequently follows *t*. This has the effect of reducing the true information content, because part of the message is devoted to telling us something we know the probability structure of the English language. Hence, a channel with sufficient capacity to transmit the 26 letters of the alphabet at a certain speed with no errors actually uses more bandwidth than required. This is well illustrated by the fact that if a message is received through a channel of lower transmission capacity, say a noisy channel, and the received message is garbled, one can usually determine from a knowledge of English what the original message was. Shannon, whose work will be discussed in considerable detail later, and to whom the illustration with English is due, states that a message in English can usually be deciphered with half the letters missing.

From the foregoing, it can be concluded that if part of the message is devoted to information concerning the probability structure, and if the probability structure is already known, the message will not be transmitted efficiently. Hence, the most efficient form for the message is one in which the probability structure is not transmitted, i.e., where the parts of the message are independent or random. Only noise is of this nature. For this reason the type of message which uses transmission facilities most efficiently is of the form of random noise.

7. *Bandwidth and Time*

There is one disturbing feature in all the analyses which have been presented. Frequent mention has been made of messages confined to a certain band and of a certain duration and of channels of a certain bandwidth, available for a certain time. As has been recognized by some writers, these concepts are not precise.

A message or signal cannot be limited precisely in both time and in frequency range; a Fourier series of a finite number of terms always extends in time from minus infinity to plus infinity. This can be seen more easily by using the theorem that a signal confined within a certain bandwidth may be exactly synthesized by placing, at points $1/2W$ seconds apart on the time scale, pulses of the form $\sin 2\pi Wt/2\pi Wt$ having the same amplitude as the original signal.[6] Each of these pulses is of zero amplitude at all other points $1/2W$ seconds apart, but has a finite amplitude between such points over the entire time scale both within and without the original signal duration. On the other hand, using the Fourier integral, it is simple to show that any signal of finite duration has an infinite frequency range.

Similar considerations apply to transmission channels. Actually, to say that a channel is of a certain bandwidth, available for a certain length of time, implies that other frequencies and times may be used for other signals without interference. In an engineering sense, it is not difficult to define these limits in terms of the inherent noise already present in adjoining frequency and time bands. This is not satisfactory for a precise, generalized study, however.

The statistical view of the properties of messages and communication channels avoids these difficulties by using limiting processes in the various definitions. Thus, for example, in Shannon's definition of channel capacity it will be noted that a theoretical inspection of the system for an infinitely long time is required in order to measure the channel capacity. Under such conditions the joint uncertainty in bandwidth and time discussed by Gabor is not pertinent, and it is permissible to speak of bandwidth with certainty, since the time duration for transmission is not exactly specified and is of no particular concern for this type of discussion. This may be clarified by considering transmissions in pulse code modulation where the pulses are of the form sin $2\pi Wt/2\pi Wt$ as discussed above. The rate of transmission of messages by use of such symbols will be given by some close approximation to the modified Hartley law, although in theory the contributions of all these pulses will remain in the channel for an infinite time.

8. Wiener

Recognition of the fact that the problems in communication engineering are fundamentally statistical in nature is due largely to the work of Norbert Wiener. We shall mention only references 7 and 8 which have appeared in book form.

The recent war initiated the work on anti-aircraft fire control which required statistical analysis of electrical data at a rate far too high foi any but automatic computation. As a result, Wiener was initially involved in the problems of filtering and predicting electrical data plus perturbing noise which was available for fire control purposes. This work is described in detail in reference 8. The other work,[7] deals with the much larger questions of the unity of the problems of communication, control and statistical mechanics for both the machine and the living organism. Much of the work is outside the scope of the present paper, and only Chapter 3 of reference 7, entitled "Time Series, Information and Communication," will be dealt with here. However, it is precisely in the interrelations of scientific fields and the application of the techniques and concepts of one science to another where the principal appeal of this work lies. Unfortunately, once the reader has left the philosophical portions of the text, he is likely to find himself bogged down in a morass of mathematical complexity where the path is in no way smoothed by the extreme conciseness of the presentation.

With respect to the subject in hand, two basic topics are developed in Chapter 3 of reference 7. These are first a quantitative definition of information for a particular situation and secondly, a measure of the rate of transmission of information. It is stated at the outset that a class of phenomena included in time series and the associated apparatus dealing with time series have to deal with the recording, preservation, transmission, and use of information. The concept of the measure of information is based upon the simplest unit, namely that involved in making a choice between two simple, equally probable alternatives. Considering the case where the probability that a certain quantity shall be between x and

$x + dx$ is $f(x)dx$, then a reasonable expression for the amount of information* associated with the curve $f(x)$ is

$$\int_{-\infty}^{\infty} [\log_2 f(x)] f(x)dx$$

It is of interest to mention the background for such a choice, as briefly related by von Neumann[9] in a review of "Cybernetics." The measure of information given above is equivalent (except for sign) to the definition of entropy given in statistical mechanics. This phase of thermodynamics has had a lengthy history for the physicist and engineer. In the same way that entropy measures the degree of disorganization of a system, or is proportional to the logarithm of the number of alternatives possible for a physical system after all the macroscopic measurements have been recorded, so now the idea of the quantity of information is related to the choice or number of alternatives in the selection of a time series. In the discussion of the work of Shannon to follow, a more complete and logical derivation is made and the discussion of this point is accordingly deferred.

Wiener then considers the general problem in which a set of observations is made on a received set of messages plus noise. It is desired to ascertain, given the probability distribution of the received signals, how much information can be obtained concerning the messages alone. Although a formal solution is indicated, it is not in a useful form, and we shall refer to the work of Shannon as suggested by Wiener.

The final point of interest is the derivation of the rate of transmission of information for the case of messages and noise having known statistical properties but derived from a source with the random properties of Brownian motion. The resulting expression is similar in form to the modified Hartley law as given by Tuller and Goldman. This also is in a form not directly applicable to the problems of the engineer.

In summary, much of the basic philosophy of the theory of communication is given by Wiener, although not in a form essential to the needs of the practicing communications engineer. The more readily understandable developments are those given by Tuller, Goldman, Shannon, and others, and it may be noted that it is probably not a coincidence that these men were associated in some manner with Wiener in their work at The Massachusetts Institute of Technology. In calling attention to the economy of presentation for the subjects of information and communication, it is not intended to minimize Wiener's contributions to the field. It is better to say with von Neumann that "It is hoped that some feeling of the book's brilliancy as well as its bounds may have been conveyed in the above sketchy remarks."

---

* Wiener states that he here makes use of a communication received from J. von Neumann.

III. The Synthesis of the Theory

1. *Shannon*

We turn now to the work of Shannon, who put all the previous work on a sound mathematical basis and carried it much farther. Shannon's work has been presented in three papers. The first[10] was delivered at an IRE Symposium, November 12, 1947; this included a derivation of the modified Hartley law, based on geometrical intuition, and contained a discussion of some of the important implications of the law. The second[10] is an expanded version of the first paper, giving the proofs of the unproved theorems given in the first paper, and introducing the new concept of entropy as a measure of information. The final paper[11] presents the complete theory and represents the most advanced account published on the transmission of information.

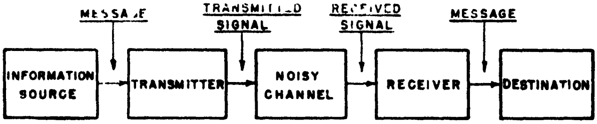


Fig. 2. The general communication situation.

It is worth while to discuss the first two papers in some detail, not only because of the new principles presented in them, but because the attack is entirely different from that of the authors previously mentioned.

A communication system consists of the various components shown in Fig. 2. The information source may be, for example, a person speaking, a phonograph record, or a scene to be televised, any of which produces a message. The message generally is a sequence of symbols or of amplitudes which follow each other in time, i.e., a time series. If the message is not already arranged in time, as for example, on the phonograph record where the arrangement is along the grooves, it is converted into a time sequence by the equipment.

There are two principal classes of messages. One class uses a finite number of symbols as message elements, code systems such as Morse or teletype being the most obvious examples. Messages in the second class are continuous functions of time, such as speech or music.

The signal is the electrical representation of the message. The essential point is that there is some definite relationship, possibly very complex, between message and signal. The signal also is a time series, i.e., a function of time. The transmitter forms the signal from the message.

The noisy channel transmits the signal, meanwhile perturbing it in a manner which has previously been discussed.

The purpose of the receiver is to reconstruct the message from the perturbed received signal.

The destination is the object or person to whom the message is directed. A complete discussion of a communication system should include the characteristics of the destination, because there is no object in designing and constructing a system to transmit messages which cannot be utilized. If, for example, the system is one designed to transmit pictures by wire for use in newspapers, the system should not be called upon to reproduce tones and resolution very much greater than can be utilized by the newspaper photo-engraving process. Because Shannon's work is perfectly general, he does not deal with the characteristics of the destination, or, for that matter, with the meaning of the message, but confines the communication problem to one of transferring a given message from source to destination. Gabor and others, on the other hand, gave considerable attention to utilizing the characteristics of human hearing to conserve bandwidth and actually designed and operated systems of adequate fidelity using extremely narrow bands.

In a manner similar to that of Goldman, Shannon shows that a signal of duration T and bandwidth W can be specified exactly by a set of $2TW$ numbers; e.g., its value at instants $1/2W$ seconds apart in time. This signal can therefore be thought of as a single point in a $2TW$-dimensional space, the $2TW$ numbers or coordinates exactly identifying the point and vice versa. Shannon points out that, since $2TW$ is a very large number for a signal of ordinary complexity, this geometrical representation amounts to using a simple entity in a complex environment to represent a complex entity in a simple environment. The simple entity in the complex environment is the signal point in the multi-dimensional space. The complex entity in the simple environment is the original signal as a function of time. The value of this representation is that, in this way, geometrical ideas can be used in the study of signals and important results can thereby be derived.

In multi-dimensional space, the "distance" from the origin to a point is the square root of the sum of the squares of the coordinates of the point, and if these coordinates are voltages, the sum of the squares is a number proportional to the total energy in the signal. Hence all signals of total energy less than E must lie inside a "sphere" of radius proportional to $\sqrt{E}$.

Since messages also can be specified by a finite set of numbers, they too can be represented as points in a multi-dimensional space. One can then think of the transmitter as a device whose function is to establish a relationship between the points of the message space and the points of the signal space. The function of the receiver is to select the message point which corresponds to the received signal point. This is equivalent to establishing a relation between the points in the signal space and the points in the message space in inverse relation to that of the transmitter. The establishment of such a correspondence is called mapping.

The mapping performed by the transmitter may be simple or it may be complex. In single sideband amplitude modulation, the signal space coordinates are proportional to the message space coordinates, and each message point is mapped into a single signal point. In broadband systems, however, the signal space is of a higher dimensionality than the message space and each message point is mapped into a set of signal points.

2. *A Geometrical Approach to the Modified Hartley Law*

When the signal is transmitted through the channel, it is perturbed by noise. This corresponds to the signal point being displaced in the signal space. If the noise is random, the displacement is equally probable in any direction, and is furthermore proportional to the square root of the noise energy, E_n. Consequently, surrounding each received signal point is a spherical region of radius $\sqrt{E_n}$, to any point of which the original signal might have corresponded. Since the energy of the received perturbed signal is equal to $(E + E_n)$, all such signals are confined to a spherical region of radius proportional to $\sqrt{E + E_n}$. Consequently, the number of signals which can reliably be distinguished is equal to the number of little spheres which can fit into the larger sphere. This is approximately equal to the ratio of the volumes of the two spheres. In n-dimensional space, the volume of a sphere is proportional to the nth power of the radius. Since $n = 2TW$ and the radii of the two spheres are proportional to $\sqrt{E + E_n}$ and $\sqrt{E_n}$, the ratio of the volumes is

$$M = \frac{(\sqrt{E + E_n})^{2TW}}{(\sqrt{E_n})^{2TW}} = \left[\frac{E + E_n}{E_n}\right]^{TW}$$

The measure of information H is the logarithm of M, the number of distinguishable signals, thus

$$H = \log M = TW \log \frac{E + E_n}{E_n}$$

where E and E_n are the signal and noise energy, respectively. The total energy is proportional to average power, hence the information is

$$H = TW \log \frac{P + P_n}{P_n}$$

where P and P_n are respectively the average signal and noise powers. Since power is proportional to the square of the rms amplitude, the information is

$$H = TW \log \left[\frac{S + N}{N}\right]^2 = 2TW \log \frac{S + N}{N}$$

where S and N are now to be understood as signal and noise rms amplitudes.

If the logarithm in the above expression is that taken to the base 2, the information capacity is given in the units of binary digits or bits. That is, in a message such as teletype, where there are only two characters, one and zero,

$$H/T = 2W \log_2 (1 + S/N)$$

is the maximum number of code characters or bits of information per second that can be sent with the given bandwidth and signal-to-noise ratio. This is not to say that any particular modulation system, such as amplitude modulation or frequency shift telegraphy, can achieve this rate. It means only that, by proper coding or choice of modulation, one can approach this limit with as small a frequency of error as desired, and that no coding system can exceed this rate of transmission under the same conditions of bandwidth and S/N ratio without a corresponding increase in the rate of error.

For messages not already in binary code the quantity $W \log_2 (1 + S/N)$, is the rate at which a message can be sent in equivalent binary digits. For example, if the message is made up from the 26 letters of the alphabet, the number n of binary digits required on the average to specify one letter or message symbol is given by

$$\begin{aligned} 2^n &= 26 \\ n &= \log_2 26 = 4.7 \end{aligned}$$

Hence the maximum rate R at which letters can be sent through such a channel is

$$R = \frac{1}{4.7} \frac{H}{T} = \frac{W}{4.7} \log_2 (1 + S/N)$$

3. Reformulation of the Problem

It is appropriate to pause at this point to consider the various concepts that have been introduced. Hartley derived a measure of the information content of a message in terms of the logarithm of the number of possible messages. He also attempted to find the maximum rate at which message symbols could be sent; he concluded, on the basis of analysis of particular networks, that this rate is proportional to the bandwidth.

Gabor refined this analysis, making it independent of particular networks, and showed that $2W$ independent symbols per second could be sent through a channel of bandwidth W.

Tuller stated that, since $(1 + S/N)$ levels of signal could reliably be distinguished through random noise for a given signal-to-noise ratio, $(1 + S/N)^{2TW}$ different messages of length T could be sent through the channel in time T and that, according to Hartley's definition, the information capacity of the channel was expressible by $2W \log (1 + S/N)$.

Shannon presented additional proof that $2W$ independent samples per second can be sent through the channel, and that $(1 + S/N)$ levels were, on the average, detectable. Both Shannon and Goldman gave useful physical interpretations of the phenomena discussed.

We are now in a position to restate the general communication problem in the light of the material already presented. It is convenient to do this in the form of questions.

1. What is a measure of the rate at which a source of messages is producing information? This measure must cover situations employing either discrete symbols such as Morse Code or continuously variable messages such as speech. It must further take account of the probability structure of the message. The measure itself will be useful if stated in the form of the equivalent rate of production of binary digits.

2. How many equivalent binary digits of information per second can be sent through a channel with a given signal power and a specified type and level of interference?

3. What coding methods can be used to transmit a message of a given information content through a channel of given capacity at the highest possible rate, especially when the message, in its original form, is of a different bandwidth from the channel?

Answers to all three of these questions are given in Shannon's principal work, "A Mathematical Theory of Communication."[11] A study of this paper can be very profitably made by considering the material pertinent to the questions presented above. We will consider in the following section the first question concerning the rate of production of information by a source.

4. *The Measure of Information*

From the previous discussion it was seen that (in the discrete case) information is produced when a selection is made from a number of possible selections. A knowledge of the quantity of information generated by a source is required for the design of an efficient channel or for choice of an efficient coding to match the source to a channel with a given capacity. In general, the source will produce symbols which are not independent; rather, a given symbol will be influenced to some degree by the preceding symbols.

Since the communication systems covered by this theory are designed to deal with classes of messages rather than individual messages, only ergodic sources will be considered. Without attempting a rigorous definition of an ergodic source, an appreciation of its properties may be obtained from a realization that any sequence generated from such a source is the same in statistical properties as any other sequence when these sequences become very long. Further, a shift of the time reference does not alter the statistical properties of any one sequence; this is generally referred to as the stationary property of a time series. These considerations infer that all long sequences have equal probability distributions and that conclusions concerning all possible sequences from a source may be drawn from an examination of one typical sequence.

Consider first a source where the information is generated by choosing successive symbols from a set of possible independent symbols with probabilities $p_1, p_2 \cdots p_n$. Then H, the measure of information, will evidently be a function of these probabilities. That is,

$$H = H(p_1, p_2 \cdots p_n)$$

An explicit form for the function H is obtained by Shannon as a result of assuming the following very general and reasonable properties for H.

1. H shall be continuous in the p_i.

2. If all the p_i for the s possible symbols are equal, then $p_i = 1/s$ and H shall increase monotonically with s. This is quite reasonable since there is more uncertainty (and hence more information) in a choice from $s + 1$ equally likely symbols than in a choice from s equally likely symbols.

3. If a selection of symbols is broken down into two successive selections, the information represented by the original choice shall equal the weighted sum of the measures of information of the multiple selection.

The only function which has these properties has the form

$$H = -K \sum_i p_i \log_2 p_i$$

in which K is a proportionality constant. Setting K equal to unity gives H in bits per symbol of the signal or message sequence. The reader is referred to Appendix 2 of reference 10 for the mathematical details inasmuch as several other deductions of this basic relation will be discussed.

Considerable insight may be gained by an alternative proof after Laemmel[12] in which the necessity for restriction of the theory to ergodic sources is much more apparent. Using the basic definition of information contained in a sequence chosen from M equally possible sequences as originally derived by Hartley, one finds

$$H' = \log M$$

where H' is the total information given by the sequence.

On the basis that a sequence of length n is made up of n different symbols, the number of possible sequences is

$$M = n^n$$

and the information per sequence is

$$H' = \log n^n = n \log n$$

It is important to note that the ergodic hypothesis infers that the various sequences become equally probable for sequences of sufficient length. Now, in general, sequences of length n will contain fewer than n different symbols, and these will generally occur with different frequencies. In fact, the number of occurrences of the ith symbol will be

$$n_i = p_i n$$

This repetition of symbols, which were originally considered to be different reduces the amount of information, since fewer sequences are available from which to choose. The reduction of information due to the repetition of the ith symbol may be written

$$H_i'{}_{\text{lost}} = n_i \log n_i = p_i n \log p_i n$$

Summing this information over all repeated symbols and subtracting from the original value yields the true information content of a typical long sequence:

$$\begin{aligned} H' &= n \log n - \Sigma p_i n \log p_i n \\ &= n \log n - \Sigma p_i n \log p_i - \Sigma p_i n \log n \\ &= n \log n - n \log n \Sigma p_i - \Sigma n p_i \log p_i \end{aligned}$$

Since $\Sigma p_i = 1$,

$$H' = -n \Sigma p_i \log p_i$$

The average information in bits per symbol may then be obtained by dividing by the length of the sequence n and using the logarithm to the base 2. Thus,

$$H = -\Sigma p_i \log p_i \text{ bits/symbol}$$

The form of H above is similar to the defining equation for entropy in statistical mechanics as discussed in the section on Wiener. The term *entropy* has been adopted in information theory for the measure of uncertainty or information.

The maximum value for the entropy with respect to the probabilities p_i subject to the condition $\Sigma p_i = 1$ is found, by application of the calculus of variations, to be attained when all the possible symbols are equally probable. This is intuitively evident since, for this condition, one is most uncertain about the outcome of any particular selection. As an extreme example, suppose one symbol to have a probability of one and all others to have zero probability. Application of the formula gives H as zero as we would expect, since there is no uncertainty about the selection.

From the preceding derivation for the form of H a somewhat more general expression may be presented. Again considering all sequences of length n for n very large, the ergodic nature of the source insures that the probabilities for all the sequences approach the same value, p. The entropy in bits per sequence is therefore $-\Sigma p \log p = -\log p$. The average entropy per symbol is obtained by dividing by the length of the sequence, whence

$$H = \frac{\log (1/p)}{n} \text{ bits/symbol}$$

From the method of derivation of the last expression for the entropy, we note that no requirement was made concerning the independence of the symbols in the various sequences with respect to the symbols preceding them. The previous definitions of entropy of a source have assumed that each symbol was independent of preceding symbols. Extending the definition of entropy to cover the uncertainty in a joint event, it is clear that the uncertainty of an event is decreased by knowledge of a past event to which it is statistically related. On this basis then, one would expect to find that if statistical relations exist over the sequence of symbols, the entropy will be less than for an equal number of symbols considered to be independent.

The difficulties inherent in considering the various possible relations between symbols are avoided by altering the scale of examination so that now the equally probable long sequences themselves are used. We have now a means for obtaining a measure of H for any given source. The process requires a statistical examination of long sequences of symbols. Two approximation expressions for the entropy are given by Shannon[11] for the usual case when the sequences can be examined over only a limited number of symbols. It is clear, however, that if the statistical relations between symbols extends over at most n symbols, then examination of sequences of greater length gives the correct value for the entropy. The better approximation according to Shannon is obtained by the following procedure.

Let H_n be the entropy of a sequence of n symbols. Consider the selection of the $(j + 1)$th symbol, denoted by k, taking into account the sequence of the preceding j symbols. The knowledge of some or all of these j symbols decreases the uncertainty in the selection of the $(j + 1)$th symbol on account of the statistical relations. Let the probability of the choice of k as the $(j + 1)$th symbol following the sequence of j symbols be called the conditional probability $p_j(k)$. The entropy H_n may then be written as the conditional entropy of the $(j + 1)$th symbol knowing the preceding j symbols. Symbolically, this may be written

$$H_n = -\sum_j p(j) \sum_k p_j(k) \log p_j(k)$$

The following example from Laemmel[12] ascribed to B. McMillan illustrates the approximation process. The entropy of sequences of various lengths of English letters is computed to be

Letters taken singly at random	4 70 bits letter
Letters taken singly with English frequencies	4 15 bits letter
Letters taken in pairs with English frequencies	3 57 bits/letter
Groups of 8 letters with English frequencies	2 35 bits/letter

The limiting value is estimated to be close to that calculated for groups of eight symbols.

An expression for the entropy of a source useful in calculating the number of important sequences from an ergodic source is

$$\lim_{n \to \infty} \frac{\log n(q)}{n} = H$$

in which $n(q)$ is the number of sequences which must be selected to obtain a total probability q after all sequences have been put in order of decreasing probability. The proof of this theorem follows readily when $n(q)$ is written explicitly in terms of q and p, i.e., $n(q) \cong q/p$. Then, since p will vary with n while q is constant, it is seen that

$$\lim_{n \to \infty} \frac{\log n(q)}{n} = \lim_{n \to \infty} \frac{\log q}{n} + \lim_{n \to \infty} \frac{\log (1/p)}{n} = H$$

From the last expressions, which are actually theorems, we may consider that an ergodic source with an entropy H can produce 2^{Hn} sequences each with a probability 2^{-Hn}. The probability of any other sequence approaches zero as the length of the sequence increases.

5. *Coding for the Discrete Source and Noiseless Channel**

Attention in this section will be directed principally toward the problem of encoding the discrete sequence of symbols generated by the information source such that maximum use is made of the transmission channel. The operations performed by the transmitter and receiver are, in general, inverse operations where the receiver operation is such as to recover the original input to the transmitter. In this case the encoding operation produces an output symbol, or sequence of symbols, for a particular input symbol. The output symbol may depend upon the state of the transmitter. The encoding operation is such as to permit a unique recovery of the input symbol by the decoding process. The entropy of the output per unit time can be shown to be equal to that of the input, indicating quite reasonably that no information has been lost or gained by the encoding process.

a. Entropy and the Ordering of Messages. It will be instructive, in view of the intimate relation between the essential information or entropy of a source and the coding process, to consider an alternate derivation for the entropy of a source in bits per symbol for the case where the symbols are not equally probable but are still independent. Fano[13] derived the expression

$$H = -\sum_{i=1}^{n} p_i \log p_i$$

as the average amount of information in bits per symbol by the following procedure. Consider the ensemble of all messages consisting of n independent selections of symbols. For each message, we seek the number of times a choice must be made between two equally likely possibilities in order to indicate or isolate that particular message. This gives the number of bits for that message. Then the average information per symbol in bits is the number of these fundamental or elementary selections divided by the length of the sequence, n. Accordingly, the procedure is to arrange the messages in order of decreasing probability and then divide the series into two groups of equal probability. The selection of the group containing the desired message is a selection between two equally probable choices. The process of division into two equally probable groups is now repeated for the group which has the desired message and, again, for that subgroup with the message chosen. This process is continued until the particular message has been separated out and the number of elementary selections required represents the information in bits for that message.

---

* The material from Shannon[11] on this topic is supplemented by the work of Laemmel[12] and Fano.[13]

It is recognized that it will generally not be possible to divide the series of messages into two groups of equal probability and, similarly, for the subgroups. However, as the length of the messages increases indefinitely, the number of possible messages increases and hence the probability of each individual message becomes smaller and smaller.

(1) Message	(2) $P(i)$ Probability of Message	(3) Div. 1	(4) Div. 2	(5) Div. 3	(6) Div. 4	(7) Div. 5	(8) Div. 6	(9) No. of bits per Message $B(i)$	(10) Recoded Message
00	0.49	0.49						1	0
		0.51							
01	0.14			0.14				3	100
				0.14					
10	0.14		0.28					3	101
			0.23						
02	0.07				0.07			4	1100
					0.07				
20	0.07			0.14				4	1101
				0.09					
11	0.04				0.04			4	1110
					0.05				
12	0.02					0.02		5	11110
						0.03			
21	0.02						0.02	6	111110
							0.01		
22	0.01							6	111111
								36	

FIG. 3. Analysis and coding of a group of messages.

The accuracy of the above process is then improved to any required extent by the limiting process except when the subgroups contain only a few messages after many divisions, although even here, the error is made small by the ordering process.

The following example, taken from Fano,[13] illustrates the above process. Consider the possible symbols 0, 1, and 2 and their respective probabilities as $p(0) = 0.7$, $p(1) = 0.2$ and $p(2) = 0.1$. Let the message length consist of two symbols. The set of messages arranged in order of decreasing probability is shown in column 1 of Fig. 3. The probability for each message is given in column 2 and the series of divisions of the group is shown in columns 3 to 8, inclusive. Referring to Fig. 3, it is noted that the number of divisions or bits in column 9 required to select the high-probability messages is less than that required for the low-probability messages, in agreement with the basic concept that there is little information conveyed by a message of high probability, and that the most uncertain message conveys most information. It is the unexpected which is most surprising! The average number of bits per symbol is obtained by adding the total number of bits for all the messages and dividing by the total number of symbols in these messages. The result is 2 bits per symbol. From the expression for H above, we have

$$H = -(0.7 \log_2 0.7 + 0.2 \log_2 0.2 + 0.1 \log_2 0.1) = 1.157$$

This is the limiting value of the entropy in bits per symbol for messages of increasing length composed of the same three independent symbols.

For messages of sufficient length, if $P(i)$ is the probability of the ith message, the number of binary selections, $B(i)$, required to specify this message is given by $-\log_2 P(i)$. This follows from the fact that $P(i)$ is the probability of the last subgroup obtained by successive division of the total probability (which is unity) $B(i)$ times so that $P(i) = 1/2B(i)$.

The average value of $B(i)$, or the number of binary selections required on the average to select one message, is given by $\Sigma P(i)\ B(i)$ where the summation is over all possible messages. The average information in bits per symbol is then

$$H = \lim_{n \to \infty} \frac{1}{n} \sum P(i) B(i) = \lim_{n \to \infty} \frac{1}{n} \sum P(i)(-\log P(i))$$

This is reducible for large n to the form $H = -\sum_{i=1}^{s} p_i \log p_i$, where this summation is over the set of symbols which may be selected.

b. The Fundamental Theorem. The stress upon the concept of H as representing the average information per symbol in bits per symbol will be justified in the following, where it will be shown that H determines the channel capacity when the coding employed is most efficient. That is, there is an irreducible quantity to be transmitted which is determined by the information contained in the message as calculated by the expression given for H. The original message, or that message formed by inefficient coding of the original, is larger than this minimum as a result of redundancy in expression, unequal probabilities for the various symbols, and the statistics inherent in any natural language. It is seen that these are the considerations which were discussed briefly above in the section *The Maximized Information Function.* The fundamental theorem for a noiseless channel, as given by Shannon, relates the entropy of the source H to the channel capacity. This theorem states that *the output of a source producing information at the rate of H bits per symbol can be encoded and transmitted without error through a channel of capacity C bits per second at the rate of C/H symbols per second.*

c. The Coding Process. Two very general ways of encoding are described by Shannon, the second of which is stated to be substantially the method of Fano partially described above. After arranging the various messages of length n in order of decreasing probability, the process of repeated division into two groups of equal probability is followed until each final subgroup contains only one message. The information contained in each message was shown above to be equal to the number of elementary divisions $B(i)$ required to isolate the message. The number of bits per symbol, H, is obtained by dividing $B(i)$ by the number of symbols, n. The process of encoding which will be described permits each message to be represented by a binary number consisting of a number of binary digits or bits equal to the number of divisions which isolated that message, and hence equal to the informational content of the message. The coding process thus encodes the message in such a way as to transfer to the channel on the average H bits per input symbol. Inasmuch as the channel capacity is assumed to be C bits per second, the rate of transmission is therefore C/H symbols per second. This rate cannot be exceeded, since the entropy of the channel input was stated to be equal to the entropy of the source as a consequence of the condition of uniqueness of recovery of the signal by the inverse of the coding process.

The process of encoding results in a representation of each message by a binary code expansion of the total probability of all the messages preceding it after arrangement in order of decreasing probability. The expansion is carried out to $B(i)$ places where $B(i)$ either denotes the number of elementary divisions referred to above or the integer equal to $-\log P(i)$ (or that integer just greater than but nearest $-\log P(i)$ if the latter is not an integer). An example of such coding is given in column 10 of Fig. 3 for the messages and probabilities listed there. It is noted that the high-probability messages, which convey least information as indicated by columns 2 and 9, may be represented by the shortest codes and, conversely, longer codes are used to transmit messages of low probability carrying most information. Since the cumulative probability of the messages preceding a particular message differs for each, the codes for each message will differ, and hence the inverse process will give a unique result. Referring to the coding in Fig. 3, it is noted that the longer codes are different from the shorter codes containing m binary digits in length in their first m digits and, accordingly, it is always possible to identify the various codes. For example, a typical sequence received and the decoded message are shown in Fig. 4.

The actual coding process in binary numbers follows the repeated division process. In the first division of the ordered series into two groups of equal probability, the code for messages in the first group starts with 0 and that for the messages in the second group starts with 1. Each group is then subdivided into two subgroups of equal total probability and the second digit of the code for messages in these subgroups is either 0 or 1 according to whether these messages fall into the first or second subgroup for each division. The process of division and assignment of succeeding binary digits is continued until each subgroup contains just one message. This is the process which gave the codes of column 10 in Fig. 3. It is noted that the 0's and 1's are used with equal frequency in this process, thus maximizing the entropy per channel symbol.

Coded Sequence	1 0 1	0	1 1 0 1	1 1 1 1 0	1 0 0	0	1 0 0	1 1 0 0
Message	10	00	20	12	01	00	01	02

FIG. 4 Illustration of efficient coding

d. The Coding Delay. The above encoding and decoding processes implicitly require examination of the sequence of symbols of the message source for coding purposes and examination of the channel coded sequence at the receiver for decoding. As a result, there is a delay in the transmitter and another delay in the receiver requiring devices in both places for storage of the messages. Practical considerations generally require that these delays be finite, whereas ideal coding is possible only when all the messages that may be transmitted are examined in the repeated division process described above. How does this examination of the transmission for a finite period affect the efficiency of coding? It has been assumed that the message sources have been statistically stationary, hence the probabilities for the various symbols can be determined by examination of the various messages for indefinitely long sequences. Examination of finite sequences for coding purposes must result in some inefficiency, since the frequency of occurrence of the symbols will fluctuate for such sub-sequences. It was shown by Shannon that the inefficiency for such finite coding delays is not greater than $1/n$ plus the difference between the true entropy and the entropy calculated for the finite sequence of n symbols. Therefore, as the value of n increases, or as longer sequences are examined for coding, the entropy for the finite sequence approaches the true entropy and the upper limit of coding inefficiency decreases. However, the actual decrease in inefficiency need not be monotonic, as pointed out by Fano.

The introduction of delay as an engineering variable is clearly emphasized by Laemmel.[12] It is quite probable that specific application of the basic theory to evaluation or design of a particular communication system will find the coding delays of equal importance with such other variables as power or bandwidth in effecting the final compromise that represents optimum design. An example which illustrates this thesis might involve a large-scale computer which gathers information from various sources, then performs calculations in very short time and finally directs the operations of high-speed controls. Efficiency in the incoming information and output control channels might introduce coding delays which would be disastrous in their effects upon the stability or accuracy of the device being controlled.

6. Coding for the Discrete Source and Noisy Channel

a. The Effect of Noise. Several new concepts of importance must be considered when the effects of noise in the channel are examined. Here the transmitted signals are perturbed by random disturbances such that the signals at the receiver are changed in some irregular manner so that it is not possible to reconstruct the message with certainty solely by operation upon the received signals. A measure of the information lost is given by the uncertainty of reconstruction of the message; this is called the *equivocation* by Shannon. The equivocation is quantitatively defined as the entropy or uncertainty of the input when the output is known, with units in bits per symbol or per second as previously used for entropy. Therefore, the information actually transmitted is given by the difference between H, the information actually put into the channel by the transmitter, and the equivocation or information loss.

This definition and usage is justified by the following reasoning. Assume an observer or auxiliary device which compares the received signals with those transmitted. This observer then uses a subsidiary transmission channel to send the information necessary for correction of the errors of the principal transmission. It may then be shown that the capacity of this subsidiary correction channel is determined by the equivocation.

As an example, consider a source producing 1000 selections per second from two equally likely symbols. Suppose that the noise perturbations during transmission are so great that the received symbols are independent of the transmitted symbols. Apparently, half of the received message will be correct due to chance alone. One might erroneously infer that the system is transmitting at the rate of 500 bits per second. A knowledge of the output still leaves the uncertainty or entropy of the input at one bit per symbol since we are totally uncertain of the transmitted signals. Hence the equivocation is 1000 bits per second and the calculated rate of transmission is zero bits per second as the result of subtraction of the equivocation from the entropy of the input to the channel. This is quite reasonable since, as Shannon notes, "equally good transmission would be obtained by dispensing with the channel entirely and flipping a coin at the receiving point."

b. The Fundamental Theorem. The fundamental theorem relating to a noisy channel of capacity C bits per second and a discrete source of entropy or rate of generation of information of H bits per second, states that if H is equal to or less than C there exists a coding system such that transmission can be accomplished with an arbitrarily small frequency of errors. If the entropy of the source is greater than that of the channel, it is possible to encode so as to reduce the equivocation arbitrarily close to H minus C, but it is not possible to make the equivocation less than H minus C.

The proof of this theorem is not as satisfying as that for the noiseless case since the proof consists in showing only that a code must exist which will have the properties listed. However, no explicit method for achieving the coding is presented. On the contrary, Shannon notes that attempts to attain a good approximation to ideal coding are generally impractical. This result appears to be related to the difficulty of giving explicitly the means for constructing a random sequence. The fundamental theorem may be reformulated in a manner which is instructive in revealing its relation to the noiseless case. This reformulation indicates that, for messages of sufficient length through a channel of capacity C bits per second, we can distinguish reliably, on the average, C bits per second where the criterion of reliability is established as a probability of error other than zero or one.

A particularly interesting property of the coding for this case is the use of redundancy in the sequence of signals to combat the effects of noise. If a piece of information is sent once, then noise may obscure the signal, but if the information is sent several times or in different forms, the probability for correct recovery of the message is increased. This explains why radar signals and television pictures may still be recognized through noise of relatively greater magnitude. The exact relationship between the various parameters of importance will be further discussed in connection with channel capacity. The problem of delay as a result of the coding process arises here as well as in the noiseless case discussed previously.

It has been shown that there is a very close relation between the entropy of the source, the channel capacity and the coding problem.

The discussion of coding for the continuous source is accordingly deferred, and will be combined with the discussion of channel capacity.

7. *The Capacity of a Communication Channel*

Perhaps the most important measure of the relative value of a communication system is the rate of transmission of information which can be attained. It is necessary, of course, that this transmission be either free from errors or that it remain within some specified limit of error frequency. These considerations have been discussed for the discrete channel and will be further amplified below.

a. The Discrete Noiseless Channel. It may seem somewhat strange to define and discuss the capacity of a noiseless channel inasmuch as such a channel is capable of an infinite rate of transmission of information. This paradox is resolved when it is realized that the channel must be considered in relation to the source of information and the transmitter. In most systems these are constrained to generate symbols chosen from a finite set and having a particular duration. In addition, there may be added constraints on the types of sequences which may be generated. In television, for example, blanking sequences are not permitted to follow each other, and in fact are constrained rigidly in their time relations Accordingly, this interrelation of source and channel is recognized by Shannon in his definition of channel capacity for the noiseless case as the limit of the time average of the logarithm of the number of *allowed* signals, as the time duration of the signals increases indefinitely. That is,

$$C = \lim_{T \to \infty} \frac{\log N(T)}{T}$$

where $N(T)$ is the number of allowed signals of duration T. For the noiseless case, it is tacitly assumed that there need be no error in the reconstruction of the original message at the receiver.

b. The Discrete Channel with Noise. It has been shown in Section 6 that the rate of transmission of information under the condition described is given by the entropy or measure of the informational input to the channel diminished by the equivocation, where the latter is the measure of the loss of information due to the effects of the noise. The necessity of considering both source and channel is again evident in the definition of channel capacity as the maximum rate of transmission with respect to all possible information sources used as input to the channel. This capacity may be achieved with an arbitrarily small frequency of error by proper coding. Many questions remain, and in particular it may be asked what, if anything, limits the rate of transmission in this case. How do the engineering parameters of signal power, channel bandwidth and noise power affect these considerations? Shannon's work assures satisfactory answers to these questions, and the following discussion will include these points.

c. The Continuous Source and Noisy Channel. In this section we shall consider the case where the signals and messages are continuously variable as in speech or commercial broadcasting. It is again assumed that the discussion will be limited to sequences which are stationary in their statistical nature. A further limitation follows from the characteristic that all communication channels and sources exhibit, namely, that the signal which is generated as a function of time is limited in its frequency band. Such functions may be determined completely by giving their ordinates at a series of discrete points spaced by a time inversely proportional to the bandwidth. Thus if the signal is limited to the frequency band from zero to W cycles per second, it may be specified by $2W$ numbers per second according to the relation

$$f(t) = \sum_{n=-\infty}^{\infty} f\left(\frac{n}{2W}\right) \frac{\sin \pi(2Wt - n)}{\pi(2Wt - n)}$$

As already noted in the discussion of bandwidth and time, the requirement that no frequency greater than W be present infers that $f(t)$ have non-zero amplitudes over the entire time scale. Nevertheless, in this representation a signal can be very closely limited to a time duration T if the values of $f(t) = f(n/2W)$ outside this interval are zero. The tails of the remaining $\sin x/x$ functions extend indefinitely in time but permit utilization of the channel because they are zero at all points spaced $1/2W$ seconds from their central maximum. We note that time limited signals from ideal continuous sources require an infinite number of coordinates for their specification where each coordinate may vary continuously. However, all real sources are limited in bandwidth, and hence the signals which are generated require only a finite number of coordinates for signals of finite duration, although each coordinate may again vary continuously. Each signal sequence of limited duration T and band W may be associated with its corresponding point in the space of $2TW$ dimensions. One may also associate each such point with the probability for this signal, thereby defining a probability distribution function p of the $2TW$ coordinates.

The measure of information or entropy of the ensemble of permitted signals may then be defined in a manner exactly analagous to the discrete case by the relation

$$H = -\int p(x) \log p(x) dx$$

where the integration is over the entire volume of the many-dimensional space representing all the possible signals. The properties of the entropy in this case are similar to those for the discrete case. The units are such that $2WH$ represents the entropy in bits per second for this distribution.

If $p(x)$ is a one-dimensional distribution, then the form of $p(x)$ which causes the entropy to be a maximum, subject to the condition that the standard deviation of x is fixed at σ, is a Gaussian distribution. This result is quickly achieved through the use of Lagrange's undetermined multipliers where H is maximized subject to the obvious condition $\int p(x)dx = 1$ and the definition $\sigma^2 = \int x^2 p(x)dx$. It is found that $p(x) = (1/\sigma\sqrt{2\pi}) \exp(-x^2/2\sigma^2)$ and when this is substituted in the expression for H, the result is $H_{\max} = \log_e \sigma\sqrt{2\pi e}$. White thermal noise possesses a Gaussian distribution for its amplitudes. The noise power N is given by the square of the standard deviation of the amplitude, and therefore by the above theorem, the signal which has the maximum entropy for a given average power, is white noise. Hence $H = \log_e \sqrt{2\pi Ne}$, and the entropy per second is $2WH$ or $W \log_e 2\pi eN$. Similar results are obtained for a multi-dimensional distribution but these will not be discussed.

The measure of information having now been defined, the question of the rate of transmission of information may again be considered. In exactly the way followed for the discrete noisy channel, this is defined as the difference between the entropy of the input to the continuous channel diminished by the equivocation or loss of information due to the noise. Similarly, the channel capacity C is defined as the maximum rate of transmission with respect to all input sources. Just as in the discrete case, the capacity is the maximum number of bits per second that can be transmitted with arbitrarily small equivocation.

Alternatively, if the message and noise components of the received signals are independent, it may be shown that the rate of transmission is given by the total entropy of the received signals less the entropy due to the noise components. The maximum value of this expression for the rate of transmission is achieved when the received signals have a Gaussian amplitude distribution or resemble white noise. This, therefore, is the case which defines the channel capacity. Let the average signal power be P and the average perturbing noise power be N. Then the received signals have a power $P + N$. The entropy of the received signals per second is then $W \log 2\pi e(P + N)$, the entropy of the perturbing noise portion is $W \log 2\pi eN$, and the channel capacity is given by the difference

$$\begin{aligned} C &= W \log_e 2\pi e(P + N) - W \log_e 2\pi eN \\ &= W \log_2 \frac{P + N}{N} \text{ bits per second .} \end{aligned}$$

This is the modified Hartley law rigorously derived in a way which sets out very clearly its theoretical and practical bases. To attain the maximum rate of transmission, the signals must possess the statistical properties of noise. As a possible coding system, a set of standardized samples of white noise is selected for the signals and groups of the message symbols are related to noise samples. At the receiver the perturbed signals are correlated with the standard noise samples and the standard sample is selected which differs least in rms value from the received signal. In general, any possible interference will have less correlation with noise samples than with non-noise signals. In fact, as the length of the noise samples increases, their correlation with any signal other than themselves approaches zero.

IV. APPLICATIONS TO TELEVISION

In the rapidly expanding field of television communication there are many more applicants for stations than there is bandwidth space available. For this and other reasons a television transmission system employing a narrower bandwidth than that of the present system would be highly desirable. The principal reason for the present bandwidth will be explained and the inefficiencies of the system discussed on the basis of the principles developed in this paper. A few alternative schemes which avoid some of these deficiencies will be presented to illustrate the theory.

In the present television system an image of the scene is focused on the plate in a camera tube. A small scanning aperture moves horizontally at constant speed in parallel lines across this image. At the receiver the picture tube screen is scanned thirty times a second in synchronism with the moving aperture at the transmitter. The electron beam intensity is made proportional to the average intensity of that part of the image momentarily being scanned in the camera tube. If there is motion in the original scene, successive images will be slightly different, for example in intensity and position of the objects in motion. The persistence of human vision produces the illusion of a constantly illuminated, continuously changing image.

The actual message transmitted consists of two parts: the signal to control the electron beam intensity, called the video signal, and synchronizing signals which keep the electron beam in step with the scanning aperture. The synchronizing signals in the present system consist of pulses at the beginning of each line and frame. The audio signal is transmitted on a separate carrier, the frequency of which is separated by a suitable amount from the picture carrier, hence it need not enter further into the discussion.

The maximum vertical resolution obtainable in this type of scanning is limited by the number of horizontal lines from which the picture is constructed, since, obviously, details smaller than the line spacing cannot be reproduced. It is desirable to provide approximately equal horizontal resolution. That is, as the electron beam rapidly moves along a line, it must be able to change from maximum to minimum intensity in the time it travels a distance equal to the line spacing. This is necessary, for example, to reproduce the vertical edge of a black object on a white background. It is well known that the bandwidth required to transmit such a change is of the order of the reciprocal of the time allowed for the change. For example, if the change must be made in $\frac{1}{5}$ microsecond, a bandwidth from $2\frac{1}{2}$ to 5 megacycles is needed, depending on the filter characteristic.

From this discussion it can be seen that the bandwidth necessary for picture transmission is completely fixed by the fact that the intensity of the beam must be changed in a given length of time. Although the sharp changes necessitating this frequency band are present only in a very small portion of most pictures, the provision of such a bandwidth permits the transmission of pictures of exceedingly complicated character; for example, a very fine checkerboard pattern wherein the side of each square is equal to the line width and the brightness of each square is completely unrelated to other squares. Furthermore, the thirty successive pictures transmitted per second may be completely different without increasing the required bandwidth. Actually, in normal television transmission, successive pictures are very much alike, except when changing from one scene to another. As a matter of fact, the pictures are required to be enough alike so that the illusion of continuous motion is produced. In addition, the illumination of the various parts of the picture must be related to each other if the picture is to be meaningful. Adjoining areas, either along a line or on successive lines, are, with high probability, of equal or almost equal intensity.

To put this in information theory terms, one has a bandwidth, W, capable of transmitting $2W$ independent data per second, but the actual data transmitted are not independent. Successive data are related; data separated by integral multiples of line duration are related; and data separated by integral multiples of frame duration are related. Consequently, the bandwidth is used inefficiently, or put another way, the information function (the video signal) is not maximized. This inefficiency from a transmission standpoint is partly compensated by other features of the present system which permit the receiver to be relatively simple in design and construction. Any schemes for improving the transmission efficiency must therefore be considered for their effect in that respect. Obviously, since there are few transmitters and many receivers, the burden of complicated equipment should be borne by the former.

The process of efficiently coding a television message would involve examination of the sequence of electrical signals put out by the pickup device and elimination of the redundancies in the message so that only essential information would be transmitted. Such examination involves storage of the signals for a time at least equal to the time taken to generate the length of the sequence being examined. A very real difficulty arises here in the matter of storage inasmuch as the rate of generation of information is so high. However, quite recently, a cathode-ray storage tube has been developed which is capable of storing a complete television picture in the form of electric charges on an insulating surface.[14] It may be noted that one other form of storage already exists in the pickup tube, where the image of the scene is focused upon a photosensitive mosaic for a time equal to the frame scanning time. An electrical image of the scene is formed by the charges on the mosaic which is then destroyed by the scanning. This type of storage is not available at the receiver, nor does a method yet exist for utilizing it at the transmitter for coding purposes.

A rather large fraction of the present channel bandwidth is required to transmit the picture frames at a rate sufficiently high so that the human eye perceives a flicker-free picture. This frame rate is considerably higher than that required for the illusion of continuity of motion and apparently contributes little to the overall transmission of information. It has accordingly been proposed that a form of storage tube be used at the receiver to store the pictures which are transmitted at a rate high enough for the illusion of a continuity of motion. The stored picture may then be scanned and presented to the viewer at a flicker-free rate. An alternative type of scanning, known as *balayage cavalier*,[15] or knight's move scan, takes advantage of certain characteristics of the human eye, just as the vocoder takes advantage of certain characteristics of the ear.

The eye cannot perceive rapid change and fine detail at the same time. This means that when a picture is changing, it is not necessary to reproduce immediately all the details of the changing scene in order to produce the illusion of continuity.

The present television system uses this phenomenon to some extent by what is called interlaced scanning. On each vertical traverse of the image, only half of the horizontal lines are scanned. There are thirty complete pictures transmitted, therefore sixty half pictures or fields, per second. This has the double effect of eliminating flicker and augmenting the impression of continuity.

In "knight's move" scanning, the principle is carried much farther by dividing the entire picture into a large number of squares, each square containing sixteen picture elements. All the squares use the same numbered pattern, each number designating a picture element. The numbering system employed follows the permissible moves of the knight in a game of chess. The scanning aperture first visits all the one's, then all the two's, etc., and eventually, after sixteen vertical traverses, the entire picture is reproduced. The author claims that this need only be done six times per second to achieve complete freedom from flicker and to create the proper illusion of continuity. Orthodox scanning methods require thirty (twenty-five in Europe) complete pictures each second. A five- or sixfold reduction in required bandwidth is therefore indicated.

It may be noted that in these examples the influence of the destination upon the transmission efficiency problem is recognized. This is not an essential characteristic of any transmission since, once the receiver has obtained the transmitted information, it is possible to present this information in any form or at any frequency without reference to or effect upon the transmission channel.

The similarity between successive pictures, and in particular, the fact that the differences are usually confined to a small area, is the basis for a suggested partial-area transmission system.[16] This coding method consists of placing the image to be transmitted on a storage screen and scanning only a part of the picture at a time. This smaller area would contain the central subject of the scene, and its detail and behavior would be emphasized by this process.

These proposals and others found in the literature are coarse approximations to the ideal coding for maximum efficiency. The extremely high rate of transmission of video information makes it quite impractical to make a complete statistical examination for coding of the message. Several other objections are opposed to these procedures in addition to the increase of receiver complexity. In the first place, the very large investment in existing television equipment makes discussion, such as that above, appear of academic interest only. Secondly, the elimination of redundancy in the course of bandwidth reduction would have a serious effect upon the service area of reception. The fringe areas are by definition those places where the signal-to-noise ratio is a limiting factor for reasonably fair reception. Adequate reception through noise interference has been counted as one of the assets of redundancy. Apparently then, the elimination of redundancy would contract the area of good reception. Alternatively, the relation of interest may be studied in the modified Hartley law where it may be noted that a decrease in W causes a decrease in channel capacity. In ordinary broadcasting, the signal power decreases as some function of the distance from the transmitter thus reducing the effective capacity with distance of transmission. To offset the contraction of service area that would be caused by the reduction in redundancy, it is necessary to increase the transmitted power.

ACKNOWLEDGMENT

The writers wish to express their thanks to the management of Sylvania Electric Products Incorporated for its encouragement and permission to publish this paper, to Mr. G. D. O'Neill for completely editing the manuscript, and to Mr. D. G. O'Connor for writing the section on *The Measure of Information* and for his indispensable assistance in coping with the multifarious demands of this project.

REFERENCES

1. Hartley, R. V. L. *Bell System Tech. J.*, **7**, 535 563 (1928)
2. Gabor, D. *J. Instn. Elect. Engrs.*, Part III, 93, 429 (1946).
3. Tuller, W. G. Sc. D. Thesis, M.I.T. (1947). A revision of this appeared in *Proc. Inst. Radio Engrs.*, **37**, 468 (1949).
4. Valley, G. E., and Wallman, H. "Vacuum Tube Amplifiers," Vol. 18, M.I.T. Radiation Laboratory Series. McGraw-Hill, New York, 1948.
5. Goldman, S. *Proc. Inst. Radio Engrs.*, **36**, 584 (1948).
6. Shannon, C. E. *Proc. Inst. Radio Engrs.*, **37**, 10 (1949).
7. Wiener, N. Cybernetics. John Wiley and Sons, New York, 1948.
8. Wiener, N. The Interpolation, Extrapolation and Smoothing of Stationary Time Series. John Wiley and Sons, New York, 1949. Originally published as an NDRC report (1942).
9. von Neumann, J. *Physics Today*, **2**, 33 (1949).
10. Shannon, C. E. The Transmission of Information. Unpublished.
11. Shannon, C. E. *Bell System Tech. J.*, **27**, 379, 623 (1948).
12. Laemmel, A. E. Report R-208-49. P.I.B. 152. Microwave Research Institute, Polytechnic Institute of Brooklyn, 1949.
13. Fano, R. M. Technical Report No. 65, The Research Laboratory of Electronics, M.I.T., 1949.
14. Pensak, L. *R.C.A. Rev.*, **10**, 59 73 (1949).
15. Toulon, P. *L'Onde Élect.*, **28**, 412 (1948).
16. *Electronics*, **21**, No. 12, 67 (1948).

Author Index

Numbers in parentheses are reference numbers. They are included to assist in locating references in which the authors names are not mentioned in the text. Numbers in italics refer to the page on which the reference is listed in the bibliography at the end of each article.

***Example*: Alberts-Shoenberg, E 214(16), *219*, indicates that this author's article is reference 16 on page 214 and is listed on page 219.**

A

Adams, N. I., Jr., 89, 97, *144*
Ahearn, A. J., 4, *42*
Alberts-Shoenberg, E., 214(16), *219*
Allis, W. P., 145, 151, 163, *181*
Alma, G., 191(8), *194*
Alpert, D., 49, *82*
Anderson, P. A., 27, *42*
Armstrong, E. H., 222
Arsenjewa-Heil, A., 52(14), *82*
Ashworth, F., 5(27), 11, 19, 27(27), 31(27), 35(27), 36(27), *42*

B

Bardeen, J., 11, *42*, 193(13), *194*
Beck, A. H. W., 44(3), 74(3), *81*
Becker, G. A, 190(6), *194*
Becker, J. A., 4, *42*
Benjamin, M., 8, 9, 11, 12, 17, 18, 19, 29, 39, *42*
Bennet, W. R., *260*
Bermer, J., 58(26), 62(29), 72(35), *82*, *83*
Biguenet, C., 191(7), *194*
Black, H. S., *260*
Bloch, F., *181*
Bode, H. W., 276, 283, *303*
Bondley, R. J., 192(10), *194*
Bott, R., 293, 294, *303*
Bradford, C. I., 210(12), *219*
Brattain, W. H., 193(13), *194*
Brillouin, L., 87(5), 90(5), 98(6), 104(8), 144(11), *144*, 148(1), 149(1), 152(1), 164(1, 2), *181*
Britten, L. F., 191(9), 194
Brüche, E., 1(1, 2), 4, *41*, *42*
Brune, O., 267(3), *303*
Brunetti, W. R., 198(4), 201(7), 204(7), 213(4), *219*
Bunting, E. N., 211(14), *219*

C

Cauer, W., 266(2), 276, *303*
Chodorow, M. I., 9(16), *42*, 53(18), 72(36), 79(37), *82*, *83*
Clavier, A. G, 255(15), *260*
Creamer, A. S., 211(14), *219*
Cuming, W. R., 197(3), *219*
Curtis, R. W., 198(4), 213(4), *219*
Curtiss, L. F., 193(12), *194*

D

Danzin, A., 192(11), *194*
Darlaston, A. J. H., 191(9), *194*
Darlington, S., 271, 281(8), *303*
Davies, J. W., 190(3), 192(3), *194*
de Boer, J. H., 16, *42*
Deloraine, E. M., *260*
Despois, E., 192(11), *194*
De Voe, C. F., 11(20), *42*
Dite, W., 255(15), *260*
Duffin, R. J., 293, 294, *303*
Dushman, S., 27, *42*

E

Edson, J. O., *260*

F

Fano, R. M., 329, 330, 332, 333, *343*
Fauve, C., 58(28), 65(30), *82*, *83*
Fay, Samuel, 89, *144*

Feenberg, E., 54, 58, 62, 65(23), *82*
Feldman, C. B., 232(12), *260*
Feldman, D., 58(25), *82*
Field, L. M., 53(17), *82*
Foster, R. M., 265(1), *303*
Fremlin, J. H., *82*
Friedman, M., 217(20), *219*

G

Gabor, D., 309, 310, 311, 312, 318, 321, 324, *343*
Gardiner, H. W. B., 190(3), 192(3), *194*
Gent, A. W., *82*
Gewertz, C. M., 275, *303*
Ginzton, E. L., 53(18), 72(33), 79(37), *82*, *83*
Goldman, S., 313, 314, 319, 321, 324, *343*
Goldsmith, H. A., 214(16), *219*
Gomm, W. H., 190(3), 192(3), *194*
Goodall, W. M., *260*
Green, N. H., 183(1), *194*
Grieg, D. D., *260*
Guénard, P. R., 44(4a), 54, 58 65(30, 31), 71(31), *81*, *82*, *83*
Guillemin, E. A., 263, 266

H

Haefer, R. H., 10, 17, 18, 19, *42*
Hahn, W. C., 44, 58, *82*
Hamilton, D. R., 44(1), *81*
Hansen, W. W., 44(4), 49, *81*, *82*
Harris, W. A., 190(5), *194*
Harrison, A. E, 44(2), 72(33), *81*, *83*
Harrison, J. S., 191(9), *194*
Hartley, R. V. L., 307, 308, 309, 311, 312, 313, 314, 315, 316, 318, 323, 329, *343*
Hartree, D. R., 86(9), 89, 108(9), 110, 111, *144*, 145, *181*
Heil, O., 49(14), *82*
Helm, R., 53(17), *82*
Henry, R. L., 205(10), 207(10), *219*
Herring, C., 2, *41*

J

Jenkins, R. O., 8, 9, 11(19), 12, 17, 18, 19, 29, 39, *42*
Johannsen, H., 1(1), *41*
Johnson, R. P., 1(1), 2, 3, 4(10), 18, *41*, *42*

K

Khouri, A. S., 201(7), 204(7), *219*
Knipp, J. K., 44(1), *81*
Kuper, J. B. H., 44(1), *81*

L

Laemmel, A. E., 326, 328, 329, 334, *343*
Lee, Ruben, 216(18), *219*
Llewellyn, F. B., 89, *144*

M

McCauley, C. E., 217(20), *219*
McNeight, S. A., 210(12), *219*
Mahl, H., 1(2), 4, *41*, *42*
Manning, M. F., 9(16), *42*
Mano, C., 191(7), *194*
Marks, B. H., 211(15), *219*
Martin, S. L., 191(9), *194*
Martin, S. T., 1(1), 8(13), 15, 18, 19, *41*, *42*
Meacham, L. A., 225, 232(9), *260*
Meinheir, C. E., 204(8), *219*
Mendenhall, C. E., 11(20), *42*
Metcalf, G. F., 44, *82*
Miller, A. R., 16(25), 36(25), *42*
Morton, J. A., 190(4), *194*
Moullin, E. B., 89, *144*
Muller, E. W., 5, 7(12), 8, 11, 12, 17, 40, 41, *42*

N

Neilsen, I., 79(37), *83*
Nelson, R. B., 4(10), *42*
von Neumann, J., 319, *343*
Nichols, M. H., 1(1), 2, 4, 11(8), 12, 15, *41*, *42*
Nordheim, L. W., 10(18), *42*
Nottingham, W. B., 27, *42*

O

Oliver, B. M., *260*

P

Page, Leigh, 89, 97, *144*
Panter, P. F., 255(15), *260*

Pensak, L., 341(14), *343*
Peterson, E., 225, 232(9), *260*
Petrie, D. P. R., *82*
Pierce, I. R., *260*
Pierce, J. R., 52, 72(32), *82*, *83*
Pohl, J., 1(2), *41*
Power, D. W., 190(5), *194*
Prakke, F., 191(8), *194*

R

Ramo, S., 58, *82*
Reeves, H. A., 222(1), *260*
Richter, G., *42*
Richtmyer, R. D., 44(7), *82*
Roberts, J. K., 16, 36, *42*
Robinson, C. S., 4, *42*
Rose, G. M., 190(5), *194*

S

Seal, R., 205(10), 207(10), *219*
Schmidt, R. W., 3(5), *41*
Sears, R. W., 237(10), 238(10), *260*
Shannon, C. E., 226, 235(14), 236(14), 256, *260*, 306, 316, 317(6), 319, 320, 321, 324, 325, 327, 328, 329, 332, 333, 334, 335, 336, 337, *343*
Shapiro, G., 205(10), 207(10), *219*
Shelton, G. R., 211(14), *219*
Shepherd, W. G., 72(34), *83*
Shockley, W., 1(1), 2, 3, 18, *41*, 89, *144*
Slater, J. C., 89, *144*, 163, *181*
Smoluchowski, R., 11, *42*
Snyder, C. L., 214(16), *219*
Snyder, W., 204(8), *219*
Sonkin, S., 79(37), *83*
Sorg, H. E., 190(6), *194*
Spangenberg, K. R., 53(17), *82*
Stoner, E. C., 86(9), 89, 108(9), 110, *144*, 145, *181*
Stranski, I. N., 12, 14, 16, *42*
Suhrmann, R., 12, 14, 16, *42*
Sultzer, P. G., 197(2), 218(2), *219*

T

Tomlin, S. G., *82*
Toulon, P., 341(15), *343*
Tuller, W. G., 204(9), *219*, 311, 312, 313, 315, 319, 324, *343*

V

Valley, G. E. 312(4), *343*
Varian, R. H., 44, *82*
Varian, S. F., 44, *82*
Victoreen, J. A., 186(2), *194*

W

Wallis, P. J., *82*
Wallman, H., 312(4), *343*
Wang, C. C., 53(19), *82*
Warnecke, R. R., 44(4a), 58, 62, 65(30, 31), 71(31), 81, *82*, *83*
Webster, D. L., 44, 50(9, 10), *82*
Weller, B. L., 210(12), *219*
White, A. B., 4(10), *42*
Wiener, N., 318, 319, 327, *343*
Wigner, E., 11, *42*
Wolfson, H., 191(9), *194*

Y

Yerzley, F. L., 1(1), *41*

Subject Index

A

Admittance
  electronic, 72
  of R C linear networks, 266
  of R L linear networks, 266
Adsorption
  binding energies, 26
  copper on tungsten, field emission study, 27ff.
  metal surfaces, 16, 20, 21
  parameters, 21ff.
  preferential, 20
  processes, 20ff.
Amplifiers
  klystron (see under Klystron amplifiers)
  pentode cascade network, 299
  subminiature audio frequency, 201
  subminiature intermediate frequency, 204ff.
  subminiature radio frequency, 199
Amplitude
  communication noise, 244
  communication noise threshold, 235
  filters, 229, 231ff.
  modulation, 224, 234, 307, 309, 310, 314, 322
  modulation receivers, 201, 224
  quantization, 229ff., 239
  quantizer, 230
  sampling, 226ff.
  signal, 249, 311, 313
  signal fidelity, 230
Anode
  fabrication materials, 191
  power dissipation in electron tube, 187
  voltage of cylindrical magnetron, 110ff., 176ff.
  voltage of plant magnetron, 111ff., 130
Atomic
  adsorption by tungsten, 7
  diameters of certain metals, 21ff.
  migration of barium, 8
  migration of copper, 30
  migration on tungsten surfaces, 7, 9

B

Barium
  atomic migration, 8
  crystal growth, 7
Barium titanate
  curie point, 211
  subminiature capacitor, 211
Batteries, subminiature, 217, 218
Binary code transmission rate, 323, 324
Binary number system, 234ff.
Bode's method, linear network input impedance, 276
Brune process
  illustrative examples, 290ff.
  network synthesis, 267ff.

C

Capacitors, subminiature, 209ff.
  barium titanate, 211
  ceramic, 205, 210
  glass, 210
  mica, 209, 210
  teflon, 210
  temperature coefficients, 211
  vitreous enamel, 210
Capacity, transmission channel, 236ff., 332, 335ff., 338, 342
Carrier waves
  continuous, 224
  pulsed, 224
Cathode ray tube, 341
Cathodes
  anti, cylindrical magnetron, 155, 157ff., 162
  anti, plane magnetron, 94, 97, 162
  emission current density, 186
  evaporation of emission coating, 187
  coating, 188, 191

  hearing aid tubes, 186
  quasi-virtual, cylindrical magnetron, 155, 157ff.
  thoriated, 4
  virtual, cylindrical magnetron, 148, 162
  virtual, plane magnetron, 90, 92, 94, 162
  virtual, vacuum tube, 135
Cauer's process
  illustrative example, 296ff.
  network synthesis, 276ff.
Cavitation, plane magnetron, 135, 140
Cavity resonators, 44
Ceramic
  capacitors, 205, 210
  cylinders, 207
  dielectric constant, 210, 212
  electrolysis, 192
  electron tube envelope, 191
  metal vacuum seal, 191
  subminiature counter, 202
  wire insulation, 215
Circuits, printed, 198ff.
  conductive paint, 213
  processes, 198, 201
Circuits, subminiature
  broadcast receivers, 201
  component assembly methods, 199ff.
  components, 197, 199, 208ff.
  cooling methods, 198, 218
  counters, 202ff.
  handie-talkie, 201
  hearing aid amplifiers, 201
  hermetic sealing, 197, 200
  insulating materials, 208
  intermediate frequency amplifiers, 204ff.
  life expectancy, 209
  molded, 199
  operating temperatures, 197
  plastic embedment, 200
  power supplies, 218
  printed, 198
  proximity fuzes, 204
  subminiaturization definition, 196
  subminiaturization techniques, 195ff.
Communication noise
  amplitudes, 244, 247
  energy, 322
  power, 222, 224, 250ff., 313, 323, 337, 338
  pulse position uncertainty, 315
  quantization, 230, 231ff.
  statistical structure, 252ff.
  threshold amplitude, 235
  transmission channel, 221, 235, 247, 252ff., 256, 311, 321, 334, 335
  white noise, 236, 247, 253, 258
Communication signal
  amplifiers, 222
  amplitude, 311, 313
  attenuation, 222
  bandwidth, 222, 229, 311, 313, 317
  effective duration, 311
  encoding, 233ff.
  energy, 322
  geometrical representation, 321ff.
  mapping, 322
  noise level, 313
  noise ratio, 222ff., 232, 236, 251ff., 258ff., 312, 314ff., 323
  power, 221, 223, 229, 232, 235, 243, 250ff., 313, 323, 336, 338, 342
  sampling, 226ff.
  sampling device, 227
Communication symbol, 307, 320
  energy dissipation, 308
  information relation, 309, 310
  interference, 308, 310
  maximized information function, 316
  occurrence frequency, 326
  transmission rate, 308
Communication systems
  coded, 316
  ergodic sources, 325, 327ff.
  frequency bandwidth, 226, 308ff., 337
  information measure, 307, 311ff., 318, 322, 324ff., 331, 337
  information measure properties, 325
  information source, 320
  information transmission rate, 309, 312ff., 318ff., 338
  message encoding, 329, 332ff.
  message encoding delay, 333ff.
  message encoding efficiency, 333, 335
  transmission efficiency, 316ff., 324, 329
  transmitter function, 320
  uncoded, 315
Complex variable
  conditions for positive real functions, 263
  positive real functions, 264

  rational functions, 262ff.
  tests for positive real functions, 263ff.
Copper
  adsorption on tungsten, 30
  field emission image, 27
  migration on tungsten, 30
Copper oxide field emission image, 32
Copper phthalocyanine field emission image, 27
  molecule structure, 27
Copper-tungsten field emission image, 40
Counters, subminiature, 202ff.
  binary assembly, 202
  decade assembly, 203
Crystals, metal
  first order reflection electron energy, 9
  growth of barium, 7
  lattice direction indices, 8ff.
  lattice parameters, 21ff.
  lattice structure, 21ff.
  structure geometry, 13
  surface adsorption, 16ff., 20, 21
  surface binding energy, 14, 20, 26
  surface emission site population density, 14
  surface structure, 13ff.
  surface tangential energy, 38
  work functions, 10ff., 15, 17, 35
  x-ray transmission, 8

D

Darlington process
  illustrative examples, 292, 295
  network synthesis, 271
Dielectric constant
  ceramic bodies, 210, 212
  temperature dependence, 212
  titanium dioxide, 211
  vitreous enamel, 210
Dissociation of diatomic molecules, 36ff.

E

Electrolysis
  ceramic material, 192
  glass, 187, 191
Electron beam
  coupling coefficients in klystrons, 55ff.
  focusing, 52
  formation, 52
  loading conductance in klystrons, 56ff.
Electron bunching
  klystrons, 45, 64ff.
  magnetron, plane, 90, 134, 141
Electron current
  cathode emission density, 186
  density in cylindrical magnetron, 149ff.
  density in plane magnetron, 90, 142
  density in subminiature tube, 186
  in cylindrical magnetron, 110, 147ff., 150, 164, 174, 176
  in cylindrical magnetron at cut-off voltage, 146, 160
  in plane magnetron, 86, 89ff., 100ff., 114ff., 150
  negative, 87, 111, 115, 125, 130
Electron gun
  design, 52
  pulse coding tube, 238ff.
Electron microscope, 1 (see also Field emission microscope)
Electron motion in magneto-electric fields, 91ff., 99ff., 148ff.
Electron trajectories (see under Trajectories, electron)
Electron transit time
  magnetron, 100, 121, 141
  subminiature tube, 190
Electron tubes (see also Tubes, electron)
  envelopes (see under Envelopes, electron tube)
  grids (see under Grids, electron tubes)
  pulse coding, 238ff.
  subminiature, 183ff.
Encoding
  efficiency, 333, 335
  group, 330
  message, 329, 332ff.
  message delay, 333ff.
  pulse code modulation transmitter, 237
Energy
  communication noise, 322
  communication signal, 322
  communication symbol, 308
  reflection electron of metal crystals, 9
  surface binding, 14, 20, 26
  surface tangential, 38
Entropy
  communication source, 327ff., 332, 335
  message, 329, 331, 333, 335

  pulse code modulation transition probability, 256
  received signal, 338
  statistical mechanics, 319, 327
  thermal "white" noise, 338
  transmission channel input, 332, 334ff.
Envelopes, electron tube
  materials, 191
  sealing, 188ff., 192
  subminiature power dissipation, 187
  temperature, 187

F

Fidelity
  measure, 232, 250 (see also Signal-noise ratio)
  pulse code modulation transmission, 247ff.
  signal amplitude, 230
Field emission from metal surfaces, 11ff.
Field emission images
  copper, 27
  copper oxide, 32
  copper phthalocyanine, 40
  copper-tungsten, 28, 29
  intensity distribution, 9, 11, 12
  tungsten, 3, 6
Field emission microscope
  cylindrical form, 2
  development, 2ff.
  magnifications, 4, 5, 38ff.
  resolving power, 3, 5, 35ff.
  spherical form, 5ff., 27
  as vacuum gauge, 33
Field emission microscopy, 1ff.
  specimen preparation, 5, 9, 10
  surface field intensity, 15
  techniques, 10, 27, 31
Filters
  amplitude, 225
  attenuation, 312
  constant resistance, 281, 297
  low pass, 312
  pass band, 313
  time, 225
  transfer characteristic, 312ff.
Frequency
  klystron converters, 50
  klystron multipliers, 47
  Larmor precession, 88, 91, 142, 149, 150, 166
  multiplex, 225
  network, 287
  network input impedance factors, 285
  signal transmission error, 237
  variation in klystron oscillation modes, 75
Frequency bandwidth
  klystrons, 44ff., 80
  magnetrons, cylindrical 141, 147, 165, 181
  magnetrons, plane, 88, 141
  signal, 222ff., 226, 229
  signal-noise ratio relation in communication systems, 315ff.
  television transmission, 340
  transmission channel, 222, 234, 236, 312, 314, 317, 323, 336
Frequency modulation, 222, 224
  communication sets, 201
Fuzes, proximity, 204

G

Gating, 242 (see Pulse code modulation, pulse shaping)
Geiger counter, subminiature tube, 193
Gewertz' method, network input impedance determination, 275ff., 295ff
Glass
  conductive paint binder, 213
  electrolysis, 187, 191
  envelope, electron tube, 187, 189, 191
  solder, 192
  subminiature capacitors, 210
Grids, electron tube
  coatings, 190
  emission, 187
  manufacture, 188, 190
  material, 190
  quantization, 239ff.

H

Hearing aid
  amplifiers, 201ff.
  subminiature tube cathodes, 186
Heil tube, 44
Hurwitz polynomials, 264ff., 273, 277, 287, 292

Hurwitz test, positive real function of complex variable, 263, 265

I

Impedance
  analytic form, 262ff.
  complementary, 279, 281, 297
  complex function, 262ff.
  linear L C network, 265
  linear R C network, 266
  linear R L network, 266
  minimum reactive input, 278, 297
  network input, frequency factors, 285
  plane oscillating magnetron 121ff , 180
  positive real input function, 271ff., 275, 286
  reactance function, 265
  specified input, network synthesis of, 290ff.
  specified transfer, network synthesis of, 275ff., 283, 286, 295
  transfer function, 275, 278, 282, 298ff.
  two terminal-pair network, 271
Inductors, subminiature, 214
Information
  equivocation, 334ff., 338
  measure, 307, 311, 318, 322, 324, 331, 337
  transmission rate, 235, 258, 311, 331ff , 336
Insulation, electric
  materials, 209
  wire, 215

K

Klystron, 43ff.
  amplifiers (see Klystron amplifiers)
  basic forms, 45ff.
  floating drift tube, 49
  frequency converter, 50
  frequency multipliers, 47
  monotron, 49
  oscillators (see Klystron oscillators)
  reflex, 48, 71ff. (see also Klystrons, reflex)
  schematic view, 45, 48
Klystron amplifiers, 66ff., 79
  cascade, 46
  cavity gap geometry, 66
  low noise, 67
  low power, 67
  power amplifier efficiency, 70ff.
  power handling capacity, 69
  two cavity, 45
  voltage gain, 68ff.
Klystron characteristics
  beam coupling coefficient, 55ff.
  beam loading conductance. 56ff.
  complex bunching systems, 64ff.
  drift tube length, 60
  efficiency limits, 69
  frequency bandwidths, 44ff., 80
  frequency limitations, 80
  gap configuration, 53
  gap interaction theory, 53ff.
  geometry, 45ff., 48
  large signal conversion efficiency, 63
  large signal effects, 62ff.
  output gap harmonics, 45
  power limitations, 80
  space charge boundary conditions, 59
  space charge effects, 58ff.
  transverse field components of gap, 5
Klystron oscillators
  electron coupled, 48
  multi-cavity, 79
  reflex, 78ff. (see also Klystrons, reflex
  traveling wave, 79
  two cavity, 47
Klystrons, reflex, 48, 71ff.
  cavity transmission line coupling, 73
  frequency variation in modes, 75
  hysteresis, 77
  long line effects, 77
  modes, 74ff.
  non-linear reflector field, 74
  normalized efficiency load, 76
  optimum efficiency load, 75
  oscillation modes, 75ff.
  power variation in modes, 75
  typical characteristics, 78

L

Langmuir's diode formula, 94, 146, 160
Larmor precession, 88, 91, 142, 149, 150 166
Lattice
  constant resistance, 283ff.
  direction indices of crystals, 8ff.

  - parameters of metal crystals, 21ff.
  - structures of metal crystals, 21ff.

M

- Magnetrons, cylindrical, 145ff.
  - anode-cathode radius, 86, 108, 110, 164
  - anode voltage, 110ff., 176ff.
  - anti cathodes, 155, 157ff., 162
  - assumptions for steady conditions, 147ff.
  - characteristic curves, 160ff.
  - current at cut off voltage, 146, 160
  - cut off voltage, 159
  - electron currents, 147ff., 150, 164, 174, 176
  - electron trajectories (see under Trajectories, electron)
  - electron transit time, 149, 164, 179
  - end effects, 86
  - equations of motion, 150ff.
  - external circuit resistance, 173
  - frequency spectrum, 141, 147, 165, 181
  - geometry, 144
  - harmonic oscillator, 153, 167
  - initial conditions, 151, 154, 167, 175
  - instantaneous position of equilibrium, 151ff., 164, 174
  - internal reactance, 179ff.
  - internal resistance, 179ff.
  - negative resistance, 146, 165
  - non-linear oscillator, 151
  - quasi virtual cathodes, 155, 157ff.
  - single stream solution, 171ff.
  - small current, small oscillations, 153ff.
  - small current solution, 168ff.
  - small oscillations, 174ff.
  - space charge current distribution, 148
  - space charge density, 164
  - static characteristics, 146, 159ff., 164
  - structure, 85, 147
  - variable conditions, 166ff.
  - virtual cathodes, 148, 162ff.
  - voltage, 151, 163
  - with R < 2, 164
  - with R > 2, 165
- Magnetrons, plane, 85ff.
  - anode voltage, 111ff., 113, 130
  - anti cathodes, 94, 97, 162
  - cavitation, 135, 140
  - characteristic curves, 94ff.
  - current density, 90, 142
  - cut-off conditions, 90, 93, 97
  - double stream solutions, 97ff., 128ff.
  - displacement current, 99
  - efficiency, 117, 124, 135ff., 138
  - electron current, 86, 89ff., 100ff., 114ff., 150
  - electron current density, 90, 142
  - frequency, 88, 141, 147, 165
  - negative resistance, 87, 94, 116ff., 132, 135ff., 139ff., 141
  - operation with current impulse, 104ff.
  - oscillations, 99ff.
  - space charge boundary condition, 100, 102, 106, 108
  - space charge current distribution, 97ff.
  - space charge density, 102, 128ff.
  - space charge voltage distribution, 96, 98, 102
  - static characteristics, 92ff., 161
  - static double stream solution, 86, 90, 97ff.
  - static internal resistance, 116
  - static single stream solution, 86, 90, 92
  - transients, 99ff.
  - transit time, 100, 121, 141
  - virtual cathodes, 90, 92, 94, 96, 162
- Magnetrons, plane oscillating
  - characteristic impedance, 119ff., 180
  - electron trajectories, 117ff. (see also Trajectories, electron)
  - high frequency impedance, 122ff.
  - low frequency impedance, 121ff.
  - negative resistance, 123ff.
  - resonance conditions, 118, 124, 141
  - resonance oscillations with direct current, 132ff.
  - resonance oscillations without direct current, 124ff.
  - small oscillations of high frequency, 117ff.
- Message
  - encoding, 329, 332ff., 341ff.
  - encoding delay, 333ff.
  - encoding efficiency, 333, 335
  - entropy, 329ff., 331, 333, 335
  - group analysis, 330
  - group encoding, 330
  - ordering, 329ff.
  - probability, 329ff.
  - storage, 341

Mica
  minimum hole spacing, 189
  subminiature capacitors, 209
Microphonics, subminiature tubes, 185
Migration
  barium atoms, 8
  copper atoms, 30
  on tungsten surfaces, 7, 9
Modulation
  amplitude, 224, 232, 234, 307, 309, 314, 322
  angular displacement, 222, 224
  frequency, 222, 224
  phase, 224
  pulse, 222, 225, 315
  pulse-duration, 224
  pulse-position, 224
  pulse-time, 314
Modulator, pulse code tube, 222, 224, 237
Molding, subminiature electron circuits, 197, 199ff., 204
Multiplex
  frequency, 225
  time, 225

N

Networks, linear
  complementary, 280, 298, 302
  constant resistance, 284, 297ff.
  damping constant, 309
  L C impedance, 265
  lossless coupling, 275, 279, 281, 299
  one Brune cycle, admittance basis, 269
  one Brune cycle, impedance basis, 268
  power reflection coefficient, 282
  power transmission coefficient, 282
  realizable transfer function, 290
  source conversion, 280, 282
  specified input impedance, 291, 293ff.
  specified transfer impedance, 296
  R C impedance, 266
  R L admittance, 266
  two terminal-pair, 268ff., 301
Network synthesis, 261ff.
  Brune process, 267, 290ff.
  Cauer's process, 276ff., 296ff.
  constant resistance lattice method, 283ff.
  coupling network, 275, 281ff., 298
  Darlington process, 271ff., 292, 295
  ladder development of input reactance function, 286ff., 300ff.
  lumped constant, 261
  minimum phase, 285, 299
  specified input impedance, 290ff.
  specified transfer impedance, 275ff., 283, 286, 295
  two terminal-pair, 286ff.
Noise, communication (see under Communication noise)
Number systems, binary, 234ff.

O

Oscillation
  klystron space charge, 45, 64, 85ff.
  magnetron, plane, space charge, 104, 107
  modes of reflex klystron, 75ff.
  resonance of plane magnetron, 124ff., 132ff.
Oscillators
  klystron (see under Klystron oscillators)
  magnetron, cylindrical, 145ff.
  magnetron, plane, 85ff. (see also Magnetron, plane oscillating)
  subminiature triode geometry, 192ff.
Oxygen
  field emission image of dissociated molecule, 37
  mobility on metal surface, 36

P

Pentode
  cascade amplifier network, 299
  coupling network, 279
  subminiature geometry, 188
Phase
  minimum, network synthesis, 285, 299
  modulation, 224
  shift, 228
Plastic, subminiature circuit embedment, 200
Potentiometers, subminiature, 213
Potting
  compounds, 200, 204
  subminiature circuits, 200, 204
Power
  communication noise, 313, 323, 336, 338

  communication signal, 221, 223, 229, 232, 243, 251, 313, 323
  dissipation, subminiature tube anode, 187
  dissipation, subminiature tube envelope, 187
  handling capacity, klystron amplifiers, 69
  klystron amplifier efficiency, 70ff
  klystron limitations, 80
  reflection coefficient, network, 282
  subminiature circuit, 218
  transmission coefficient, network, 282
  variation in reflex klystron modes, 75
Pulse code modulation, 221ff
  coding tube schematic, 238, 240
  decoding, 245ff.
  decoding circuit, 245
  decoding, graphical representation, 246
  information transmission rate, 256ff.
  output signal-noise ratio, 254, 255
  pulse gating, 242
  pulse regeneration, 244
  pulse reshaping, 225, 243ff.
  pulse shaping, 242
  pulse slicing, 242
  receiver decoding, 244ff.
  receiver diagram, 243
  receiver operation principles, 243
  transition probabilities, 248, 249
  transition probability entropy, 256
  transmission error probability, 252
  transmission, graphical representation, 248
  transmitter diagram, 237
  transmitter encoding, 237
  transmitter holding circuit, 237
  transmitter operation principles, 237ff

Q

Quantization
  amplitude, 229ff., 236, 239
  graphical representation, 241
  grid, 239ff.
  noise, 230, 232, 247
  signal-noise ratio, 232
  steps, 229, 231
Quantizer, amplitude, 230
Quantum
  range, 231, 235
  states, 231ff., 234

R

Reactance
  internal of cylindrical magnetron, 179ff.
  network synthesis, input function, 286ff., 297
  oscillating magnetron, 122ff.
Reactors, subminiature, 216
Receivers
  amplitude modulation, 201, 224
  broadcast, subminiature circuits, 201
  pulse code modulation, decoding, 244ff.
  pulse code modulation, diagram, 243
Reciprocal function of complex variable, 264
Residue condition, 274, 278, 292, 302
Resistance
  external circuit of cylindrical magnetron, 173
  internal, of cylindrical magnetron, 179ff.
  negative, of magnetron, 87, 94, 123ff., 132, 135ff., 139ff., 141
  oscillating magnetron, 116ff., 121ff.
  steady state magnetron, 116
Resistor, subminiature, 213
Resolving power
  emission microscope, 3, 5, 35ff.
  Heisenberg limits, 40
  particle electrons, 37
  television systems, 340
  wave electrons, 39
Resonance
  condition, electron trajectories, 119, 133
  plane magnetron, 118, 124ff.
Resonator, cavity, 44

S

Sampling, signal, 226
Scanning, television
  interlaced, 341
  "knight's move," 341ff.
  storage tube, 341
Sealing
  electron tube envelope, 188ff., 192
  hermetic, of subminiature circuits, 197, 200
  vacuum, ceramic-metal, 191

Signal, communication (see under Communication signal)
Signal-noise ratio, 222ff., 232, 236, 251ff., 258ff., 312, 314ff., 323
  relation to communication system bandwidth, 315ff.
Slicing, pulse code modulation, 242
Space charge
  boundary conditions, cylindrical magnetron, 148
  boundary conditions, klystron, 59
  boundary conditions, plane magnetron, 100, 102, 106, 108
  current distribution, plane magnetron, 97ff.
  density, cylindrical magnetron, 164
  density, plane magnetron, 102, 128ff.
  distribution, plane magnetron, 104ff., 108
  effects, klystron, 58ff.
  infinite density, 102, 107
  oscillations, cylindrical magnetron, 173
  oscillations, plane magnetron, 104, 107
  reduction factor, klystron drift tube, 60ff.
  voltage distribution, plane magnetron, 96, 98, 102
Statistical mechanics, entropy, 319, 327
Stethoscope, electronic, 201
Sturm test, positive real function of complex variable, 263

T

Teflon, 199, 209, 215, 216
Teletype
  information transmission rate, 323
  signal symbols, 307
Television, 335, 339ff.
  message coding, 341, 342
  message storage, 341
  partial area transmission system, 342
  picture scanning, 341
  resolution, 340
  system, 339
  transmission bandwidth, 340
  transmission efficiency, 340
  transmitted message, 339
Temperature
  coefficient of subminiature capacitors, 211
  electron tube envelope, 187, 198
  subminiature electron circuits, 197
Thermal noise, 236, 247, 253, 258
  amplitude distribution, 338
  entropy, 338
Thyratron
  current requirements, 186
  subminiature, 187, 203, 204
Trajectories, electron
  constant anode voltage, cylindrical magnetron, 109
  contact, 102
  crossing, 87, 109ff., 115, 119, 121, 125
  exponential current impulse, plane magnetron, 105
  negative resistance, plane magnetron, 139ff.
  oscillating plane magnetron, 117ff.
  oscillating plane magnetron with direct current, 133ff.
  oscillating plane magnetron without direct current, 125ff
  rectangular current impulse, plane magnetron, 107
  resonance, plane magnetron, 133
  self-consistent, cylindrical magnetron, 150, 155ff., 166, 176
  self-consistent, plane magnetron, 86, 118ff., 150
  single stream motion, plane magnetron, 124ff., 131
  space charge limited current, plane magnetron, 101ff.
  temperature limited current, 114ff.
Transceivers, subminiature, 201
Transducer, subminiature, 193
Transformers, subminiature, 193
Transition probabilities, 247ff.
  entropy, 256
Transmission
  binary code rate, 323, 324
  coded, 314ff.
  pulse code modulation error probability, 252
  x-rays through metal crystals, 8
Transmission channel
  capacity, 236, 237, 332, 335ff., 338, 342
  carrier waves, 224
  discrete, noiseless, 336
  discrete, with noise, 336ff.

  - frequency bandwidth, 222, 234, 236, 312, 314, 317, 323, 336, 340
  - frequency error, 237
  - information capacity—bandwidth relation, 309, 311, 312
  - information rate, 235, 258, 311, 331ff., 336
  - links, 221, 225, 244, 252, 255
  - noise, 221, 252ff., 310ff., 321, 322
  - noise cleaning, 222ff., 236
  - noise effect, 334ff.
  - noise energy, 322
  - pulse code modulation rate, 256ff.
  - sideband noise level, 314
  - signal-noise ratio, bandwidth relation, 315
- Transmitter
  - function in communication systems, 320
  - klystron relay, 79
  - pulse code modulation diagram, 237
  - pulse code modulation encoding, 237
  - pulse code modulation holding circuit, 237
- Triode, electron tube
  - planar microwave geometry, 190
  - subminiature oscillator geometry, 192
  - U.H.F. subminiature geometry, 188
- Tubes, electron subminiature
  - advantages, 184
  - anode materials, 191
  - anode power dissipation, 187
  - cathode manufacturing techniques, 190
  - current density, 186
  - economics, 189
  - envelope materials, 191
  - envelope power dissipation, 187, 191
  - envelope sealing, 189, 192
  - envelope temperature, 187
  - fabrication mechanics, 188ff
  - gas cleanup, 186
  - geiger counters, 193
  - grid emission, 187
  - grids, 190
  - life expectancy, 184, 185, 192
  - limitations, 184ff.
  - microphonics, 185
  - miniaturization limits, 184
  - pentode geometry, 188
  - planar microwave geometry, 190
  - rectifier, 192
  - reliability, 185
  - solid state devices, 193
  - thyratron electrode geometry, 187
  - transducer, 193
  - triode oscillator geometry, 192, 193
  - voltage regulators, 192
- Tungsten
  - conduction electron energies, 8
  - crystals, 3ff.
  - field emission images, 3, 6, 7
  - subminiature tube filaments, 191
  - surface atomic mobility, 7, 9

V

- Vacuum
  - ceramic-metal seal, 191
  - gage, 33
- Velocity modulation tubes, 43ff.
- Vitreous enamel
  - dielectric constant, 210
  - insulation, 209
  - subminiature capacitors, 210
- Voltage
  - cut-off, cylindrical magnetron, 159
  - cylindrical magnetron, 151, 163
  - cylindrical magnetron anode, 176ff.
  - gain in klystron amplifiers, 68ff.
  - plane magnetron anode, 111ff., 113, 130
  - regulators, 192

W

- Waveforms, coding tube pulses, 239
- Wire
  - electrode insulation, 215
  - space utilization factor, 214, 215
- Work function metal crystal surfaces, 10ff., 15, 17, 35

X

- X-ray transmission through metal crystals, 8